Estimation of Population Totals from Imperfect Census, Survey and Administrative Records

Bernard Baffour-Awuah

Division of Social Statistics

School of Social Sciences

Faculty of Law, Arts and Social Sciences

University of Southampton

October, 2009

Declaration

I, **Bernard Baffour-Awuah**, declare that this thesis titled, '**Estimation of Population Totals from Imperfect Census, Survey and Administrative Records**' and the work presented in it are my own.

I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given;
- with the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ________________________________

Date: ________________________________

Dedicated to my Dad, Opoku, for instilling in me the belief that I can be whomever I want to be. Also to my Mum, Felicia, and Pseudo-Mum, Elspeth, without whose constant encouragement, support and love I most probably would not have got this far.

Acknowledgements

First and foremost, *Gloria in Excelsis Deo.*

I would like to thank my supervisors, Professor Peter W.F. Smith and Doctor James J. Brown for their help and guidance throughout this thesis.

I would also like to thank my friends and colleagues in the Division of Social Statistics, and the School of Social Sciences in general. The PhD experience will definitely have been a pretty lonely, unchallenging and unenjoyable one without them. In particular, I would like offer a special mention to my cohort, Guy Abel, Claire Bailey, David Clifford, Lisa Danquah and Alexandra Skew. Guy Abel was specifically a great help in developing the SPLUS/R programs used in this thesis. In addition, Thomas King read and provided very useful comments on various draft chapters.

My sincere thanks and gratitude to my family, both in Cape Coast and Edinburgh, for their unstinting support and love, and for putting up with my ever shifting deadlines!

Finally, this research has been made entirely possible by the Economic and Social Research Council (ESRC) who supported me through studentship number PTA-031-2004-00273.

Abstract

The theoretical framework of estimating the population totals from the Census, Survey and an Administrative Records List is based on capture-recapture methodology which has traditionally been employed for the measurement of abundance of biological populations. Under this framework, in order to estimate the unknown population total, N, an initial set of individuals is captured. Further subsequent captures are taken at later periods. The possible capture histories can be represented by the cells of a 2^r contingency table, where r is the number of captures. This contingency table will have one cell missing, corresponding to the population missed in all r captures. If this cell count can be estimated, adding this to the sum of the observed cells will yield the population size of interest. There are a number of models that may be specified based on the incomplete $(2^r - 1)$ table of observed counts, and if a model is found that adequately fits these observed counts an estimate of the unobserved cell can be derived. The thesis will be concentrating on the log-linear model specification of capture-recapture models.

In the simplest capture-recapture model, there are two lists (for example, a Census and a Survey) leading to a 2x2 contingency table, with three observed counts and an unobserved cell count. By assuming there is independence between the Census and Survey, an estimate of the unobserved cell can be obtained. It will be shown that when there is information from individual capture in the Census, Survey and a third (the Administrative List) it is possible to account for different dependencies, specifically the association between capture in the Census and Survey. The assumption of independence which is pivotal to the case when there are only two captures can now be relaxed. However, the introduction of the Administrative List means that overenumeration cannot be assumed to be negligible.

Therefore, the proposal is to use latent class models, where the idea is that there is a latent variable with two classes - one representing the real enumerations and the other, erroneous enumerations. Under the classical parameterisation of latent class models, there is the assumption of local independence, implying that the Census, Survey and Administrative List are conditionally independent given the latent variable. Consequently, when an individual's enumeration in the Census is associated with their enumeration in the Survey this latent model is invalidated. There are a number of locally dependent latent class models, but within a triple system scenario most encounter problems regarding model identifiability; to be precise, the model solutions are not unique. Thus the thesis investigates the use of the Expectation Maximization (EM) algorithm to fit a locally dependent (and identifiable) latent model to capture-recapture data from three systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

The importance of an accurate head count of the population is something that cannot be over-emphasized. Most countries in the world conduct a regular census of their populations through either a direct enumeration of the population, surveys, administrative systems or hybrid schemes that combine some or all of the aforementioned methods. These days, reliable census data are not needed at just the national level but also at a supra-national level; for example EU member states have a statutory requirement to produce periodic reliable population estimates. In the UK, the decennial census is used as a basis from which other official statistics are derived. As such its functionality is fully maximized when it achieves a near-complete coverage of the population. Paradoxically, it is inevitable that whenever a census is held, there will be some people missed, and the experiences of the 2001 UK census - and internationally in the 2000 round of censuses - suggest that there will be challenges to the achievement of high coverage of the population.

Now in the 2001 UK census, a dual system estimation methodology was employed, using an initial population count (the 'Census'), and a post-enumeration survey (the 'Census Coverage Survey'). The (Census Coverage) Survey was basically an intensive repeat enumeration of the population, for a small sample of areas. The purpose of the Survey was to assess how well the Census enumerated the population, and inform the extent of underenumeration, overenumeration and the accuracy of the responses. An underlying assumption of the dual system approach was that the Census and Survey processes were independent; when there is a lack of independence then bias is introduced into the population estimates. However, there is no way of determining the extent of this bias, unless through some additional information. Furthermore, overenumeration, which could occur through people being duplicated or erroneously enumerated, was assumed to be negligible. Thus, data from a third source - an administrative records system - is proposed in this thesis as a way of augmenting for this bias. This is the definition of triple system estimation being proposed in this thesis, and is equivalent to a three-system capture-recapture approach as employed in biological populations.

1.2 Context of Research

When both the Survey and Census have relatively low levels of coverage (as happened in some areas in the 2001 census), the dual system estimation methodology faces some difficulties in estimating the population undercount with accuracy. In recent decades many countries have encountered problems with underenumeration when conducting a traditional census, and this is becoming particularly challenging in Western countries. The traditional census is only viable when participation amongst the general population is high. Germany, for example, has not carried out a traditional census since 1987, and it has been even longer in the Netherlands which has not had one since 1971. In both these countries the previous census response rates had been very low. To that end, they have turned to alternative data sources in the form of administrative registers. It has to be said that the countries that currently rely on administrative registers have spent a remarkable amount of resources developing and maintaining them, and so under most circumstances the traditional census enumeration still remains an indispensable, and the most viable, option.

There has not been a lot of work undertaken in combining traditional census enumeration with population registers, and this is where most of the PhD research will be focused. It will be looking at how information from existing administrative registers can be used to supplement the census process. In sum, the thesis will be exploring techniques to obtain population totals when there is data from three imperfect data sources. The imperfections are because firstly, some people will be missed by all three sources, and secondly some people will be counted more than once or wrongly counted.

1.3 Objectives of Research

The main objectives of the research are two-fold:
a. to bring under one framework the existing methodology of capture-recapture methods as applied to censuses,
and
b. to provide a log-linear modelling framework that can estimate the population size, with an estimate of both the underenumeration and overenumeration.

1.4 Organisation of Thesis

The thesis first reviews the present literature on capture-recapture and population measurement and then goes on to present the different methods used to estimate the population size in general multiple capture-recapture models. Data from capture-recapture experiments can be represented in a contingency table, and the relationships that exist between the cells can be investigated by log-linear modelling. The focus of the thesis is on the three capture-recapture model since the overarching aim is to investigate how population totals - with an adjustment for the level of underenumeration and overenumeration - can be derived when there is data from the initial census enumeration, a post-enumeration survey and an administrative list. There has been a wide amount of literature that deals with adjusting population totals for underenumeration; however the literature that examines methods for adjusting for overenumeration, on the other hand, is sparse. In actuality, the early capture-recapture models made an explicit assumption that all individuals had been correctly identified, which basically translated to mean that there is no overenumeration.

In the thesis the proposed method of accounting for the level of overenumeration is through latent class analysis, and as such there is a detailed review of latent class modelling, in particular within the framework of capture-recapture models. As will be explained in Chapter 3, by writing the latent class model as a log-linear model the existing capture-recapture models can be extended to cope with overenumeration. The interpretation of the unobservable latent variable is that it is responsible for the observed patterns and associations in the contingency table counts, and can be thought to represent an underlying classification in terms of the enumeration status. Here, the general idea is that the latent variable is made up of two distinct latent classes - one characterizing the real enumerations and the other, erroneous enumerations.

Although in capture-recapture the observed data likelihood is often intractable, the complete data likelihood is usually relatively simple. Therefore, the Expectation Maximization (EM) algorithm will be employed to maximize the likelihood and derive the (unknown) population size estimate. It will be shown that within the log-linear latent class framework the EM algorithm can be naturally extended to determine the population size, adjusted for both underenumeration and overenumeration.

The thesis is organised into six other chapters, in addition to this introductory chapter. Chapters 2 and 3 review the historical and methodological literature. Chapter 4 presents the results of a simulation study of different population estimators. Chapter 5 is a 'real' census application of the techniques proposed in the thesis. Chapter 6 outlines how population estimators could be derived for triple system data that has both overenumeration and underenumeration. Chapter 7 gives a summary of the thesis and presents some suggestions as to how these estimators could be expanded and generalised to different applications as well giving a brief sketch of some ideas for future work.

Chapter 2 provides a brief historical review of capture-recapture methods and places census measurement within this context, with a focus on the measurement of underenumeration in UK population censuses since the 1980s. Dual system estimation, with its underlying assumptions, is introduced. When these assumptions are contravened there is some bias introduced into the population estimates derived under dual system estimation. The chapter therefore concentrates on two such violations - heterogeneity and dependence - and considers some ways to ameliorate the bias that arises. One such way is to move to triple system estimation and so the next part of the chapter reflects on the existing administrative sources in the UK, and in particular Scotland. Scotland, as well as being a microcosm of the UK, does have a reasonably maintained health register, which can be used in addition to the Census and Survey in triple system estimation. However, the assumption of no overenumeration that is required under dual system estimation is still needed for the basic triple system estimation model. If the overenumeration is believed to be an underlying characteristic of individuals that cannot be directly measured but manifests itself through the observed patterns and inter-relationships, it is possible to model this through latent class models. As such the final part of the chapter expounds on latent class models, and notes that these models are actually not new in capture-recapture since they have previously been used to account for heterogeneous captures within the population. Finally it concludes with a brief review of Bayesian methods.

Chapter 3 gives more of a methodological overview of the techniques that are used in the thesis. The original ideas of capture-recapture methods were developed in wildlife population measurement, but as time has progressed the methods have found uses in a wide range of applications, and there has been an expansive amount of literature. The motivation of this chapter is to bring together this literature under one framework and in application to a triple system census. Besides reviewing the literature it also seeks to clarify concepts and definitions so as to contextualise the work undertaken in this thesis. The results that appear within this chapter are, in the most, not new since they appear in the texts cited. However, the contribution of this part of the thesis is to present a methodology that can be used to estimate the population size for data collected from three systems that have both overenumeration and underenumeration.

This methodology employed is hugely reliant on log-linear modelling and the Expectation Maximization (EM) algorithm. Initially in Chapter 3 the assumption is made that there is no overenumeration, so results of the missing cell estimates under different list dependencies (which can be modelled as a log-linear model) are presented. The latter part of the chapter, focusing on the case where there is both overenumeration and underenumeration, introduces latent class modelling. In the literature there are two parameterisations of the latent class model: one based on conditional probabilities, and referred to as the Goodman parameterisation, and the other is based on log-linear models and is referred to as the Haberman parameterisation. For both parameterisations the latent model under triple system estimation is presented in this chapter. Though closed form solutions exist when there are no erroneous enumerations, it is much more difficult to estimate latent class models due to the latent, unobservable, variable, and hence this motivates the EM algorithm. The chapter concludes by proposing a model that can be fitted using the EM algorithm that copes with both overenumeration, underenumeration and dependence.

Chapter 4 is an evaluation of different population estimators in a simulation study, while Chapter 5 presents the results of an application. In Chapter 4 the objective is to assess the performance of different dual and triple system estimators when there are differing levels of dependency introduced. The initial simulations presented are for a simulated population with dependence but no overenumeration. Later on in the chapter, results are presented for the simulation study conducted to ascertain the impact of overenumeration on the performance of the dual and triple system estimators. In Chapter 5, data was obtained from the US Census Dress Rehearsal that was carried out in 1988, prior to the 1990 US census. Different dual and triple system models were then fitted to these data to derive estimates of the missing population, and compared. Owing to the measures taken by the US Census Bureau to clean and validate the data during the Dress Rehearsal, an assumption was made that there is no overenumeration. This assumption is checked in Chapter 6.

Chapter 6 presents a generalized framework for estimating population totals when there is data from three lists that are imperfect measures of the population. Three issues are considered here, namely dealing with dependence, heterogeneity and identifiability. The EM algorithm implemented in the previous chapters is developed further to cope with both dependence and heterogeneity. Using a grouping covariate an identifiable model is fitted to some Feasibility Study data and also the US 1990 Dress Rehearsal data to give an alternative interpretation to the results in Chapter 5. Finally, measures of precision of the parameter estimates are also presented.

Chapter 2

Review of Literature

2.1 Introduction

The strength of a census rests upon its near-complete coverage of the population. However, the experiences of the 2001 One Number Census in the UK (and in support, the 2000 round of international censuses) suggest that the challenge of achieving high coverage was, and is going to be, substantially difficult. One of the issues from the analysis of the last census is that modern societal changes do impact on the census methodology. Undeniably, the 'traditional family' household that was the norm in the 1970s, and to some extent in the 1980s, is very much different to that of the new millennium. The current UK population household structure has an increasing number of cohabiting couples, families with part-resident children and multiple occupancy households. Additionally, some sections of the population are increasingly mobile. Moreover, there is some difficulty in constructing adequate address lists with clearly defined vacant, derelict, communal establishments and commercial properties in the UK. The postal address file was used in the 2001 census, and provided a starting point for enumerators, who were meant to amend them to account for hidden or new households. But this does have some challenges: based on Scotland data supplied by the Scottish Government Housing Statistics Department every day 4 new dwellings are formed by conversion of old properties, 66 dwellings by new developments and 10 dwellings are demolished[1].

The review of literature undertaken seeks to give an overview to how census underenumeration measurement has progressed. With the continuing difficulties encountered during population measurement, more innovative techniques are needed to adjust the initial census counts for underenumeration. So the review initially starts with the historical development of capture-recapture methods. It goes on to give an overview of the US and UK census underenumeration strategies. The rest of the review concentrates on the methodology of triple system estimation, firstly looking at the motivation, given that the

[1] Data supplied courtesy of Jan Young, Scottish Housing Statistics, Scottish Government and are for the period 2001-2004.

basic capture-recapture model assumes homogeneity and independence, and these were pertinent issues after the 2001 One Number Census, and secondly considers latent models in coping with overenumeration in the systems. The review also looks at the different approaches of population size estimation, both in the classical and Bayesian paradigms.

The chapter is organised as follows. Section 2.2 gives a brief historical account of capture-recapture methods in censuses and details the census undercount measurement, Section 2.3 brings into focus the UK census undercount measurement since 1981 and introduces dual system estimation while Section 2.4 explains how dual system estimates could be biased due to heterogeneity and dependence and goes on to present how an administrative list could correct this bias through triple system estimation. Section 2.5 gives a brief overview of current overenumeration measurement strategies. Since the prevailing way of presenting capture-recapture models is through contingency tables, Section 2.6 describes some methods of analyzing contingency tables, specifically when there is data from three sample captures. Section 2.7 introduces latent class models and expounds on how these models are to be used to cope with overenumeration. Finally Section 2.8 presents a review of Bayesian capture-recapture models models.

2.2 History of Capture-Recapture Methods in Censuses

The problem of estimating the size of a population is one of the oldest statistical problems. Seber (1982) traces the first use of capture-recapture back to the 18th century when Laplace sought to measure the population of France in 1786. The present statistical framework owes a great deal to the pioneering work of Petersen (1896), Lincoln (1930) and Schnabel (1938) on dual lists. In their time, the application of capture-recapture sampling was primarily intended for the estimation of ecological populations - for example, Petersen and Lincoln's work focused on estimating the size of fish and waterfowl populations residing in their natural, wild habitats.

Although the origins of the models and methods of capture-recapture estimation lie in human population measurement, it has always had a firm footing in ecology because of its intuitive appeal. The use of capture-recapture techniques in epidemiology came much later (the earliest paper being Wittes and Sidel (1968) who evaluated the frequency of birth defects). One of the main reasons for this could be the fact that the independence and homogeneity assumptions, which were pivotal in earlier work, could not be reasonably applied in an epidemiological setting. The development of the log-linear framework allowed for the generalization of the basic dual list problem to multiple lists; in the main because it allows for dependence among lists and heterogeneity of capture among individuals.

In 1949, Chandrasekar and Deming applied the approach to the estimation of the birth and death rates using a population register and a survey. This proved to be a major advance in using capture-recapture techniques in human population measurement

(Chandrasekar and Deming (1949)). Their technique, later given the name dual system estimation, found an application in census estimation when it was used to adjust the totals obtained from the census and the follow-up survey.

Notably, the US Census Bureau have sought to measure census errors using a special survey since the 1950 census. However, it was in 1980 that an explicit use of the dual system estimation methodology was employed, via data from the census and a survey (known as the Current Population Survey). This resulted in a set of estimates that subsequently could be combined to give estimates of underenumeration (Hogan (1992)). It was first noted here that the census underenumeration was non-random, and varied disproportionately by social stratum and race. Therefore, any re-enumeration survey will also be likely to be susceptible to underenumeration, and miss people in the sampled areas. Thus, the dual system methodology allowed for an adjustment to be made for different population groups; in effect underenumeration varied by geography and socio-demographic factors.

Unfortunately the 1980 census was criticised for using such an evaluation programme to adjust for underenumeration - previously any follow-up survey served the purpose of measuring the quality of the census, so the final 1980 census population estimates were left unadjusted. This is because in the United States as well as the census counts being used to distribute federal funds, they also serve as a basis for the apportionment of congressional representation. Consequently, underenumeration became politically important and inevitably, there was some questioning of the census results, leading several city and state governments taking the US Census Bureau to court (see Werker (1981)).

The initial ruling found in favour of the plaintiffs and concluded that the methodology for the 1980 census was reasonable in its objective to adjust for differential underenumeration since a failure to adjust the initial census counts would lead to an underestimation of some areas and a subsequent loss of funds in areas where the census failed to count a significant proportion of the population - a case in point was New York State, and specifically New York City. It was, however, admitted that despite the detailed procedures set out by the Census Bureau, there were some problems in enumerating the population in large cities (Hogan (1992))[2]. This does show what a tricky issue census adjustment is; as well as being statistically robust, there is the requirement for it to be 'politically robust'.

Following on from the problems after the 1980 census, statisticians began to realise that the achievement of complete-coverage was an almost impossible task (Ericksen et al. (1985)). The futility of the task was due to the fact that trying to achieve complete coverage relied on procedures that were not cost effective, and in most cases increased the number of erroneous inclusions. They advocated for a properly carried out census that achieves as high coverage as possible, followed by a well designed follow-up survey to adjust for the differential levels of undercount. A subsequent article by Ericksen et al. (1989) set out the strategies for dealing with the issues raised by the 1980 census. The

[2] This ruling was subsequently overturned on appeal by the US Census Bureau (Werker (1981)).

key assumption in the design of the follow-up survey was that the sample blocks are made up of much more homogeneous sub-sections of the population, as recommended by Chandrasekar and Deming (1949).

In 1990, a much larger scale evaluation programme was undertaken, and the language moved from achieving 'complete-coverage' to 'near-complete' with a more 'statistically defensible method of adjustment' (Hogan (1993)). The evaluation programme was based on demographic estimates and a post-enumeration survey (PES). The PES was designed in a similar manner to the previous censuses, but with a much more detailed focus on sampling. A national sample of block clusters was selected after stratifying on region, race, housing tenure as well as age and sex at the national level based on what was known about the distribution of census underenumeration in 1980[3]. Hogan (1993) concluded that operationally the 1990 post-enumeration survey was a success and was practically feasible in that it achieved the data processing in the specified time-frame. Nevertheless, as in 1980 the issue of underenumeration was subjected to considerable litigation and contestation; more so, after the then Secretary of Commerce, R.A Mosbacher, announced his decision not to adjust the 1990 census[4].

Much of the debate surrounded the ability of the dual system estimation and the models used to adjust the census counts in non-sampled areas. Furthermore, there was the assumption of homogeneity between the individual states which was deemed contentious. Thus for instance in California, the estimate of the state population was calculated using synthetic estimates of individuals in the different post-strata as found in the state but the same post-strata could have applied for an entirely different state, such as North Carolina. As such, the synthetic adjustments were found to have some failings attributed to the fact that the PES did not have a large enough sample to facilitate the production of direct estimates for the state totals (Skerry (2001)).

After the 1990 census and in the build-up to the millennial census further refinements were made to the census estimation strategy. This was because when evaluating the 1990 census, it was discovered that certain population sub-groups were more likely to be missed by the census - this is what is referred to as biased or differential undercount. Hogan (1993) claimed that the only methodology that can feasibly measure the amount of differential undercount at relatively low levels of geography was a large scale post-enumeration survey followed by dual system estimation. As a consequence, the 2000 census involved an initial census count followed by an independent coverage measurement survey similar to the 1990 census. To that end, a larger post-enumeration survey was implemented which took

[3] Hogan (1992) and Hogan (1993) provide detailed discussion of the 1990 census methodological and operational processes.

[4] Again New York City, and a number of cities with large numbers of ethnic minority residents, sued the federal government to compel for adjustment of the census. The judgement was made in favour of the defendants, upon which subsequent appeals were made. Finally in January 25, 1999 the US Supreme Court ruled that adjusted figures may not be used to apportion congressional seats but it may be permissible for other purposes, where feasible (Skerry (2001)).

account of the differential underenumeration as well as to produce small area population estimates that were fairly unbiased. The key difference was that the 2000 census post enumeration survey was much larger than that carried out in the 1990 census (i.e. 300,000 housing units in 2000 compared with 165,000 in 1990).

The work carried out by the US Census Bureau over the previous censuses has shown that the estimation of census underenumeration is entirely feasible. A properly designed large scale follow-up survey that re-enumerates a sample of small groups of housing units should be undertaken independently of the initial census enumeration. This allows the initial census counts to be adjusted for the estimated underenumeration, since the survey sample facilitates the estimation not only of those missed by the census but those incorrectly counted by the census. The dual system estimation methodology also accounts for the fact that the survey will fail to count its areas perfectly and as a consequence some individuals will be missed by both the census and the survey.

2.3 History of UK Census Undercount Measurement

The 2001 census was, to all intents and purposes, the first census in the UK to make a serious concerted effort to measure the census undercount and adjust the population estimates to correct for undercount. This is not to say that the previous censuses did not attempt to measure census errors. In fact since the 1971 census, there has been a process of evaluating both the quality and coverage of the results ((Brown, 2000, page 21)). For example, in 1971 census there was some adjustment of the population counts, by demographic analysis using the mid-year population estimates.

The next census in 1981 was, however, the first to use a separate survey, the Follow-up Survey, to assess the quality of the census. The census evaluation programme was undertaken on the basis of this survey, using a stratified multi-stage sampling design. Firstly, the UK was stratified by region and area (of which there were four types - metropolitan, non-metropolitan, inner-London and outer-London). From these strata, a sample of 300 blocks (known as enumeration districts) was selected, of which 29 were in Scotland. To account for the fact that it is advantageous to assess the coverage of the census in particularly difficult to enumerate areas, the selected blocks were graded using the national classification of residential neighbourhoods (see Webber (1977)). Thus, enumeration districts classified as difficult to enumerate were selected with probability proportional to twice its estimated size. The second stage chose a cluster of four households per selected enumeration district. In the same vein, in districts classified as difficult to count under the Webber (1977) classification, the number of chosen households was doubled. Britton and Birch (1985) state that the objectives of the 1981 evaluation programme were three-fold:

(a) to check whether all persons present on census night in a private household had actually been correctly enumerated by the census;

(b) to verify the classification by the census enumerators of unoccupied residential accommodation;

(c) to assess the quality of replies given to the census questions, and hence the accuracy of the published 1981 census results.

As the objectives show, the focus of the 1981 census evaluation programme was mainly to identify the types of census errors, i.e. the quality of the census outputs. The evaluation of the census coverage was of secondary importance because undercount was thought to be small - the net level of undercount was estimated to be 0.45% (Britton and Birch (1985)).

In 1991, a much more integrated approach was undertaken to assess the quality and coverage of the census, using the Census Validation Survey. This survey was similar to the 1981 Follow-up Survey. It involved a multi-stage sampling design, with an initial selection of enumeration districts, then an enumerator selected samples of households to assess the different sources of census errors. The Census Validation Survey of 1991 was found to be unsuccessful in assessing the census coverage (Heady et al. (1994)). This was because comparisons with the demographic estimates showed the 1991 census failed to take account of the differential levels of undercount, especially amongst young males. For example, the sex ratios of adjusted census counts for young males were found to be below one. There were two possible explanations why this might have happened. Firstly, there may have been a disproportionately larger mass of undetected emigration between 1981 and 1991 for men than women. The second, and more plausible, was that there was a differential underenumeration of young males which the Census Validation Survey was unable to detect.

The sampling design of the Census Validation Survey relied on the assumption that underenumeration was homogeneous across (suitably defined) groups of the population. Hence, the major metropolitan cities were all assumed to have the same level of underenumeration. This level of aggregation was found to be too high, because differential underenumeration existed within these cities - the less affluent parts of the metropolitan cities experienced much higher proportions of people missed. Another shortcoming of the 1991 Census Validation Survey was that it was difficult to ascertain at which level the undercount adjustments were to be made, whether at ward or enumeration district level. This brought about some difficulties as to the validity of the census counts, particularly in terms of resource allocation (Brown, 2000, page 24). Additionally, the undercount figures were only known by age and sex, so it was difficult to discern how the people missed differed from the general population by other important characteristics such as education and employment.

Obviously, these issues were instrumental in influencing the design of the 2001 census. Therefore in 2001, the population measurement strategy used a new methodology with the primary purpose of adjusting the census counts to account for this differential undercount. There was a general perception that the problems in identifying underenumeration in 1991 were due to operational difficulties. These operational difficulties were not realised in 1981; for example, enumerators encountered problems contacting households and individuals - in both the actual census and the survey. One of the main causes of this is due to changing household patterns. A lot more people now live in multi-occupied houses, and in purpose-built flats in buildings that utilise electronic entry systems. Additionally, changes in employment patterns and an increase in out-of-home activities - consequences of modern demographic behaviour referred to as the second demographic transition (as proposed by Lesthaeghe and van de Kaa (1986)) - had an adverse effect on census response. In order to get around this, the 2001 census was the first to use a postal enumeration strategy. Although it does reduce part of the problem, the removal of the crucial enumerator - respondent interaction leads to an additional complication in the identification of households.

Brown (2000) does say that compared to other similar census-taking countries, the 1991 UK census was not a particularly poor census. The specific problems encountered were in counting special population sub-groups, for example Armed Forces personnel and students. He goes on to put it succinctly, on page 25, that
"*the more serious problem* [in 1991] *was not so much the existence of the underenumeration, but the inability of the* [Census Validation Survey] *to measure underenumeration.*"

Accordingly in 2001, an independent follow-up survey (known as the Census Coverage Survey, CCS) was carried out after the official Census had taken place. Matching techniques were used to link records from the Census Coverage Survey to those from the Census. This results in a 2x2 contingency table (see Table 2.1), and the objective is to obtain an estimate of the people missed by both the Census and Census Coverage Survey. A key assumption here is that the first and second processes are (statistically) independent.

Table 2.1: Dual System Estimation

		Census Coverage Survey	
		Counted	Missed
Census	Counted	n_{11}	n_{10}
	Missed	n_{01}	n_{00}

In 2001 this independence assumption was met by post-stratification at a low level of geography (here, postcode) of the population by age, gender and other covariates (identified through the Hard-to-Count (HtC) Index, which is discussed later). When the population has been suitably stratified it can now be reasonably assumed that the probability of enumeration in the second process given enumeration in the first is identical to the probability of enumeration in the second, given that person was missed in the first process. This identity assumption provides a basis for estimating the number of people that were not enumerated in either the Census or the Survey. Consequently for each stratum, this population undercount was combined with the census population to give an estimate of the true population, adjusted for undercount, of the sampled areas.

In effect the dual system estimation process used in the 2001 census methodology estimated the people missed by both the Census and Census Coverage Survey (n_{00}) by considering the relative number of people observed by

- both the Census and Census Coverage Survey (n_{11});
- the Census but not the Census Coverage Survey (n_{10}); and
- the Census Coverage Survey but not the Census (n_{01}).

Dual system estimation relies on two assumptions - independence and homogeneity. Firstly, there is independence between the processes that yield the Census and Census Coverage Survey counts, leading to estimates that are ultimately unbiased. Secondly, for any age-sex group within a chosen postcode, the probability of a person being in the Census or Census Coverage Survey is assumed to be the same for all individuals. The first condition is met by ensuring that the Census and Census Coverage Survey processes are operationally independent. Simulation work undertaken by Brown et al. (1999) demonstrated that, provided the response rates are high, the effect of any dependence is minimal, even for extreme levels of dependence. The majority of postcodes are small and can be assumed to contain 'similar types' of people; hence the second condition is likely to be met.

In addition to the fact that the dual system estimates depend on the assumption that the events of inclusion in the Census and Survey are independent, it is also important that there is an appropriate sampling scheme which chooses blocks for inclusion in the Census Coverage Survey such that inferences from the sampled blocks reasonably extend to the unsampled blocks. The probabilities of inclusion in the Census and the Survey are known to depend on the various characteristics of the population, thus post-stratification (based on the Hard-to-Count Index) was used to produce sub-groups with relatively homogeneous inclusion probabilities.

The HtC index (see Brown et al. (1999) and Chapter 3 of Brown (2000)) was a more efficient way of grading enumeration districts. The previous censuses used the Webber (1977) classification and simply oversampled areas that were deemed to be hard to enumerate. Another shortcoming of the Webber classification was that it was based on deprivation. The 1991 census showed that although a disproportionate number of poorer people were missed, suggesting a link between underenumeration and deprivation, there were other related indicators (Brown et al. (1999)). Therefore, the HtC index utilized all the variables associated with census underenumeration, taking account of both deprivation and transiency. Areas with high numbers of privately rented, multi-occupied households and young migrants were indicative of highly mobile (or transient) populations.

Eventually, on completion of the 2001 census, estimates of the populations for each local authority by age and sex were produced using a combination of regression and small area techniques. Households and persons estimated to have been missed by the census were imputed to produce a fully adjusted Census database. All population estimates were quality assured using demographic analysis and comparisons to aggregate level administrative data. Further adjustments were made, if deemed necessary, to meet the consistency requirements. The methodology of the 2001 One Number Census project is outlined in Brown et al. (1999) and given in a lot more detail in Steele et al. (2002).

Albeit the 2001 census methodology is deemed to be the best presently available for carrying out a conventional census, it does not (unfortunately) correct for the most extreme circumstances. There was evidence of poor enumeration in some areas with address lists failing to capture substantial redevelopment (Treasury Select Committee (2002)). In areas where this was particularly severe, the ability of the census methodology to make a robust adjustment for the differential undercount was stretched. Therefore, in Manchester and Westminster City Councils there were potential discrepancies between the council administrative lists and the address lists collated by the Office for National Statistics. It must be noted that the perceived failures of the One Number Census in Manchester and Westminster were driven by separate issues. In Manchester, the Census Coverage Survey design was not detailed enough to cope with rapid regeneration, predominantly attributable to the Commonwealth Games held there in 2002. However, in Westminster there was a general failure of the enumeration process with a 74% enumeration rate, compared to a national rate of 94% (see Office for National Statistics (2004) and Statistics Commission (2004)). Thus, the Manchester and Westminster Matching Studies set out to investigate the census estimates. The results of the studies led to minor revisions of the 2004 mid-year estimates (Office for National Statistics (2004)). One point to come out of these studies is that there needs to be continuing dialogue between the census takers and the local authorities before, during and after the census, so that any concerns are rectified much earlier.

2.4 Issues Surrounding Census Underenumeration

The previous sections have given a historical overview of census methodology, concentrating on how methods have been developed in the US and the UK that seek to supplement the enumeration process and improve the initial census count using a carefully conducted and executed sample survey. This section looks at two issues surrounding dual system estimation in modern censuses - dependency and heterogeneity - and then assesses how an additional list, in the form of an administrative records list, can be used to adjust the census process in light of these issues.

2.4.1 Dependency and Heterogeneity

As mentioned briefly earlier on, the simplest capture-recapture model involves two samples and relies on five key (but untestable) assumptions:

- the population is closed,
- individuals can be matched from capture to recapture,
- the capture in the second sample is independent of capture in the first, and
- capture probabilities are homogeneous across across all individuals[5]
- there are no erroneous captures in either the first or second sample.

In the most elementary two-sample capture-recapture model, it is impossible to ascertain whether the samples are independent. Usually some control can be exercised by the experimenter in the design to ensure that the assumptions are met, nevertheless the matching, independence and homogeneity assumptions are closely intertwined. This is because homogeneity of capture is a condition for independence; and vice versa[6]. Further, the assumption of homogeneity follows from the matching assumption since the latter implies that marked and unmarked individuals have the same probability of being caught in the second sample, so that capture in the first sample does not affect capture in the second. What this shows is that a failure of any one of these assumptions can invalidate the others.

In most cases failure leads to biased population estimates. This bias is termed correlation bias and can be due to two types of dependencies:

(i) List dependence: - the act of being included in the first list makes an individual more or less likely to be included in the second list, i.e. inclusion in the first sample has a direct causal effect on inclusion in the second. This is sometimes referred to as *causal dependence.*

(ii) Heterogeneity: - even if the two lists are independent within individuals, the lists may become dependent if the capture probabilities are heterogeneous among individuals. This is similar to the Simpson Paradox which shows that an aggregation of two independent 2x2 tables may result in a dependent table. This is sometimes referred to as *apparent dependence.*

In practice, these two types of dependencies are confounded and cannot be separated unless additional information is provided. There is also the likelihood that, in human populations especially, the homogeneity of capture within lists may be violated. After the results of the 2001 One Number Census were analysed there was some concern as to the validity of one (or both) the homogeneity and independence assumption for some scenarios (Simpson et al. (2003) and Brown, Abbott and Diamond (2006)). If people in the same post-stratum have different probabilities of response, then the same people might be likely to be omitted from both the Census and Census Coverage Survey. In this instance, estimates based on the independence assumption have a correlation bias which is indicative of a systematic under-estimation of the true population.

[5] Technically, this assumption can be relaxed and Wolter (1986) demonstrated that it is only required to have homogeneity across one list.

[6] Again strictly speaking at an individual level, as long as there is independence between the list capture probabilities, then homogeneity of capture, although desirable, is not necessary.

On the other hand, the direction of the correlation bias due to the effect of causal (or list) dependence is less certain: the effect inclusion in the Census has on the individual's propensity to be included in the Census Coverage Survey can be either negative or positive. An individual included in the Census could be deemed more aware of the Census process and hence would be more likely to participate in the Census Coverage Survey, than those individuals missed by the Census. This correlation bias is positive and leads to an under-estimation of the population estimate. Alternatively, the individual could feel that they have already responded to the Census, and hence would be more resistant to be included in the Census Coverage Survey, than someone who was originally missed. This type of dependence leads to an over-estimation. It is difficult to ascertain which of these two types of causal dependence is more likely. In other words, it is unclear as to whether correlation bias due to list dependence would lead to under-estimation or over-estimation of the population under dual system estimation.

The difficulty lies in the fact that these assumptions are untestable. Thus, several authors have made significant contributions to relaxing some of them. In the context of census underenumeration, work has been undertaken that looks into relaxing the independence assumption. Isaki and Schultz (1986) proposed several alternative dual system estimates to incorporate the correlation bias due to list dependence. Wolter (1990) and Bell (1993) suggest the population totals and sex ratios as additional demographic information to assess the dependence. Zaslavsky and Wolfgang (1990) and Zaslavsky and Wolfgang (1993) proposed using an administrative list as a third system. Alho (1990) and Alho et al. (1993) modelled heterogeneity using a logistic model containing several explanatory variables under the assumption of independence. If there are no covariates that can explain the heterogeneity, then individuals can be thought of as having some random effects that determine their catchability in each sample. So under the assumption of independence across individuals, Darroch et al. (1993) and Agresti (1994) allow for heterogeneous capture probabilities by using a logit model with random effects. This model is the same as that suggested by Rasch (1960) in an application to educational testing, with individuals differing on a continuous scale.

2.4.2 Administrative Lists as a Source for Census Estimation

One simple way of getting around the restrictive assumptions imposed by two-sample capture-recapture is to increase the number of samples. This has been the preferred option in biological capture-recapture. Thus in a census application, triple system estimation, proposed by Zaslavsky and Wolfgang (1990) and Zaslavsky and Wolfgang (1993), brings data from an administrative list matched to both the census and survey. The main advantage is that a third list allows for the possibility of two-way interactions between the counts derived from the different sources. Therefore, the independence assumption which underlies dual system estimation is no longer necessary. The advent of log-linear models has also made it possible to consider a far greater number of models, under different dependency scenarios. However, it may be necessary in some cases (e.g. log-linear models) to ensure that when there are multiple recording systems, each record system has homogeneous inclusion, such that individuals have the same probability of capture (within sub-strata)[7]. So in a census context, the individuals can be captured in three possible sources - namely the Census, post-enumeration Survey and the administrative List. After matching, the data can be represented in terms of a 2x2x2 contingency table with cell counts given by $\{n_{ijk}\}$, where 1 means counted, 0 means missed, and i is the Census, j is the Survey and k is the Third List (shown in Table 2.2).

Table 2.2: Triple System Estimation

		Third List			
		Counted		Missed	
		Survey		Survey	
		Counted	Missed	Counted	Missed
Census	Counted	n_{111}	n_{101}	n_{110}	n_{100}
	Missed	n_{011}	n_{001}	n_{010}	n_{000}

Administrative lists have been widely applied in population estimation, with several Western European countries (the Netherlands, Norway and Sweden, for example) relying on population registers as their main source of population estimates. These countries use surveys to evaluate the coverage and quality of the administrative records. In the UK, population estimation from administrative records has been proposed but remains controversial, partly because of issues pertaining to privacy and confidentiality. Nevertheless, falling census participation and the improvement of administrative records due to technological advancement has led to experimentation using administrative data. Advocates cite the wide availability of administrative sources, their ease accessibility in digital form, the rapid development of new information technologies and the computationally intensive statistical methodological advances.

The use of administrative data for the production of statistics is not an entirely new concept in the UK - it could be said that this is one of the fundamental reasons for the collection of such data. An example that illustrates this point is the use of Department for Work and Pensions (DWP) records on those in receipt of unemployment benefits to calculate the level of unemployment. There are currently a wide range of administrative sources that collect data on employment, education, housing, business, etc. from the population, and this is one of the main reasons why countries have resorted to harnessing this richness in their population census estimation. In some Nordic countries that rely solely on administrative data for census estimation, there is an obligation for the national statistical office to first examine whether the data exists in an administrative source before commencing on a data collection process (UNECE (2007)). The advantage of such

[7] When there is heterogeneity introduced in this way, one solution is to use Latent Class Modelling, considered in detail later.

an approach, in addition to the obviously more efficient use of resources, is that there is minimum inconvenience to the population which can have the effect of improving response rates.

There is a plethora of potential ways in which administrative records can be used to improve census coverage. In triple system estimation two scenarios where administrative records can be employed in population estimation come to mind. Firstly, the Survey and the Administrative List could be assumed to cover the same sampling blocks. This would lead to an estimate of underenumeration using triple system estimation for the blocks where the Census, Survey and Administrative List counts are available. Standard estimation techniques can then be applied to the adjusted counts to gain estimates of the population. Under a second scenario, the Administrative List records could be assumed to be available across the whole country. These records could be matched to each other and the Census, yielding a combined list across the entire country. This augmented list of the population can then be matched, in the sampled areas, to the Survey; dual system estimation can effectively be implemented.

The assumption is that different types of people are missed by either the Census or the Administrative List, so combining them yields greater population coverage. The Survey can be designed as usual with the assumption that within the sampled blocks there is near-complete coverage. Both scenarios have their advantages and disadvantages, but according to Stuart and Zaslavsky (2002) even though the second scenario requires assembling much larger files, the logistical cost would not be proportionally more than that required to obtain an Administrative List for just the sampled blocks, since the same systems must be accessed. The issue here is that the administrative data collection methodology is very different to that of the census or survey processes. Thus the assumption of independence is now more reasonably valid. Moreover, the addition of the third source brings extra degrees of freedom allowing dependency between the census and survey processes to be taken account of during the estimation.

During the 1980s - particularly in the period leading up to the 1990 census - the US Census Bureau conducted an Administrative List Supplement programme, with the primary purpose of evaluating the feasibility of including information from administrative lists in census coverage assessment. It should be noted that the US Census Bureau use the E-list to represent those individuals enumerated by the Census, the P-list to represent individuals enumerated by the post-enumeration survey process, and the A-list to represent the individuals on the administrative list.

The culmination of this programme was the 1988 Census Dress Rehearsal where triple system estimation - in application to the census - was first trialled out. On the basis of data from the 1980 census, it was found that black males were particularly difficult to enumerate (Zaslavsky and Wolfgang (1990)). Furthermore, the post-enumeration survey did not achieve sufficiently high levels of coverage for this subgroup either. So the estimates under a simple dual system estimator were thought to yield biased population estimates, due to the likelihood of correlation bias. To that end, an inner-city area of St Louis, Missouri, was chosen owing to it having a large Black population resident in tenement buildings. The A-list was constructed from merged state and federal government drivers' licence, tax revenue, military selection and military veterans administrative records (Darroch et al. (1993)). This data will be considered in greater detail in Chapter 5.

It follows that in theory triple system estimation is superior to the dual system approach. However, in practice the needs of an administrative list bring additional complications. For example, the matching of three different lists is a non-trivial matter. Presently (in the UK and for that matter the US) there is a distinct lack of a single high quality, reliable and accurate administrative data source that encapsulates the whole population, at all levels of geography. The alternative is linking several databases, as done in the US (by Stuart and Zaslavsky (2002) and Stuart and Judson (2003)) but this is problematic in terms of the different individual record-identifiers used.

An example of such a linked database is the Statistical Administrative Records System (StARS), set up by the US Census Bureau in the run-up to the 2000 census. This consists of seven merged data sets including Internal Revenue Service returns, drivers' licensing, selective service files, Medicare records and residence information from the Department of Housing and Urban Development Tenant Resident Assistance Certification System (TRACS). TRACS is effectively similar to the UK postal address file (PAF). However, it has the advantage that it is operated by a US government department. The major problem of the StARS is that the potential for duplication on it can be high since individuals will appear on the different lists, at different times. Furthermore, the address information on the files may be incomplete, incompatible or erroneous, thus leading to geocoding errors, with people counted in the wrong place. Additionally, the files may not be completely current. For example, although Drivers' records may be kept up-to-date since a driving licence is most people's sole mode of identification in the US, Medicare records on the other hand may not be updated after a move, unless the person needs medical assistance.

2.4.3 UK Population Registers

A population register incorporates information on births, deaths and mobility within a geographical area, for instance a city or council area. Some registers can also incorporate immigration and emigration. They have the advantage that they can be updated on a regular basis, and population registers are considered the future of census taking (Martin (2007)). Many countries have some form of population register at either a local or national level that covers certain population subgroups - e.g. the Department of Work and Pensions (DWP) has information on benefit recipients and national insurance contributions. In principle, one or a combination of these registers can replace the traditional census; in practice the quality and usability of such information at very low levels of geographical detail constrains the feasibility of registers as a main source of population estimates. Unlike countries where population registers have been in use for decades, the UK has not had a single administrative database that encompasses the full spectrum of the population sub-types. In countries where this is not in existence there is at least a concerted effort to coordinate the existing recording systems, using individual-specific multi-purpose identification numbers.

In the Netherlands, for example, a Social Statistics Database is used as a population register, and is constructed by combining a variety of lists. The key to the success of this database is the existence of the unique social-fiscal number. Although there is no law making registration compulsory, generally speaking it is virtually impossible to function in Dutch society without being registered. As an example, any one working without being registered pays the top rate of income tax (Schulte Nordholt et al. (2004)). The difference between this and the UK National Insurance number is the fact that every Dutch resident uses their social-fiscal number to access key services, on a regular basis, and in most cases commits it to memory. Also, unlike the National Insurance number which is applied for when seeking employment or access to benefits, the social-fiscal number is given at birth. Obviously there is a major obstacle to such registers being fully implemented in the UK because of the public perception and acceptance of them. Key to the Dutch Social Statistic Database functioning as well as it is, is its approval by the general population[8].

In the UK, health registers are the current best candidates to serve the purposes of a population register. In their current format health record data in Scotland are the most useable when compared to England, Wales and Northern Ireland for these purposes. In Scotland, there are two health-related registers - the National Health Service Central Register (NHSCR) and the Community Health Index (CHI). The NHSCR is the oldest and was set up to organise payments to doctors after the National Health Service was set up in 1948 (although there was a fore-runner established during the 1939 pre-World War II national registration). It has been carefully maintained to contain one record for

[8] There were some demonstrations for a period in the 1970s when it was being implemented. Thus a wide scale exercise was undertaken to make the public aware of why the database was a good idea, and how it was a means of improving the ways they accessed key services.

every resident in Scotland. Later on (circa 1970) the CHI was set-up as a regional primary care pilot data repository used to manage child and adult health screening programmes, in addition to facilitating payments to general practitioners. After the success of the pilot the CHI was rolled out to other health authorities in Scotland. Unfortunately the CHI was administered by the different health authorities, and so when someone migrated to a different health board they were given a new number. The differences between the CHI and NHSCR are minor and mainly operational. The key difference being that the NHSCR contains basic demographic patient information, while the CHI contains a great deal more (see Table 2.3), and therefore is more useable as a population register.

Table 2.3: Information contained on the CHI

a.	CHI number	b. NHS Number
c.	Date of Birth	d. Sex
e.	Surname	f. Birth Surname
g.	First Forename	h. Second Forename
i.	Alternative Forename	
j.	Marital Status	k. Previous Surname
l.	Date Surname changed	
m.	Address	n. Postcode
o.	Area of Residence	
p.	Reason for transfer	
q.	GP code	
r.	GP GMC Number	
s.	Date Accepted on GP list	
t.	GP name	
u.	Practice Code	
v.	Contact Date	
w.	List of Hospital contacts	

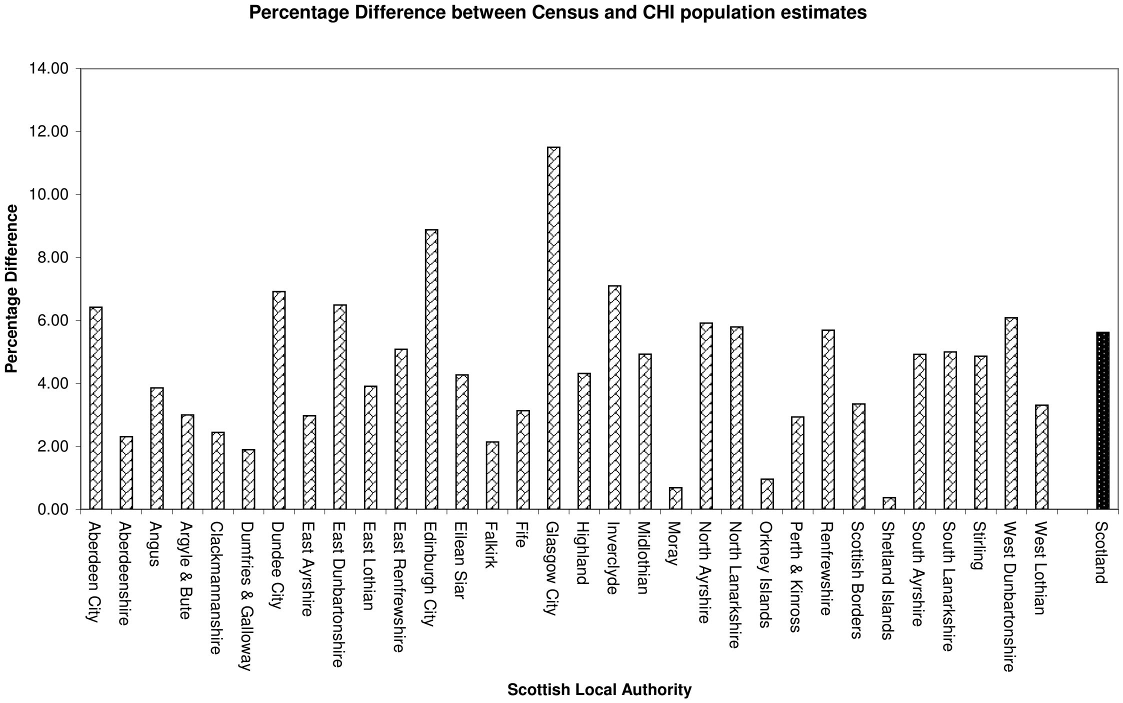
Source: Ganka Mueller, Demography Division, General Register Office for Scotland

Since its inception a person's CHI number has been a 10-digit number, of which the first six digits are their date of birth and includes a gender identifier. So, in the late 1990s work was carried out to ensure that each individual had a unique patient identifier, i.e. the CHI number. Previously, as pointed out, it was common for a person to be registered more than once on the CHI due to migration. The one-to-one correspondence was accomplished by first creating one database from the eight different CHI databases which maintained computer records about the people living in different geographical areas in Scotland. This used the fact that when a person migrates and re-registers with a new medical practice, a transfer request is made for the patient's medical records. Therefore, it was possible to use probability matching to link the patient's CHI details across health boards making duplicate records a far rarer occurrence[9]. The CHI number now currently appears on every NHS registration card, and is allocated to every to new born baby, or new GP registration. Unfortunately, one of the problems of using the CHI as a source of population data is due to the problem of list size inflation. As the data used to update the CHI comes from medical practice list records, the removal of emigrants and deaths

[9] The difficulty occurs if the person has emigrated to other parts of the UK and failed to re-register with a new GP.

tends to take some time.

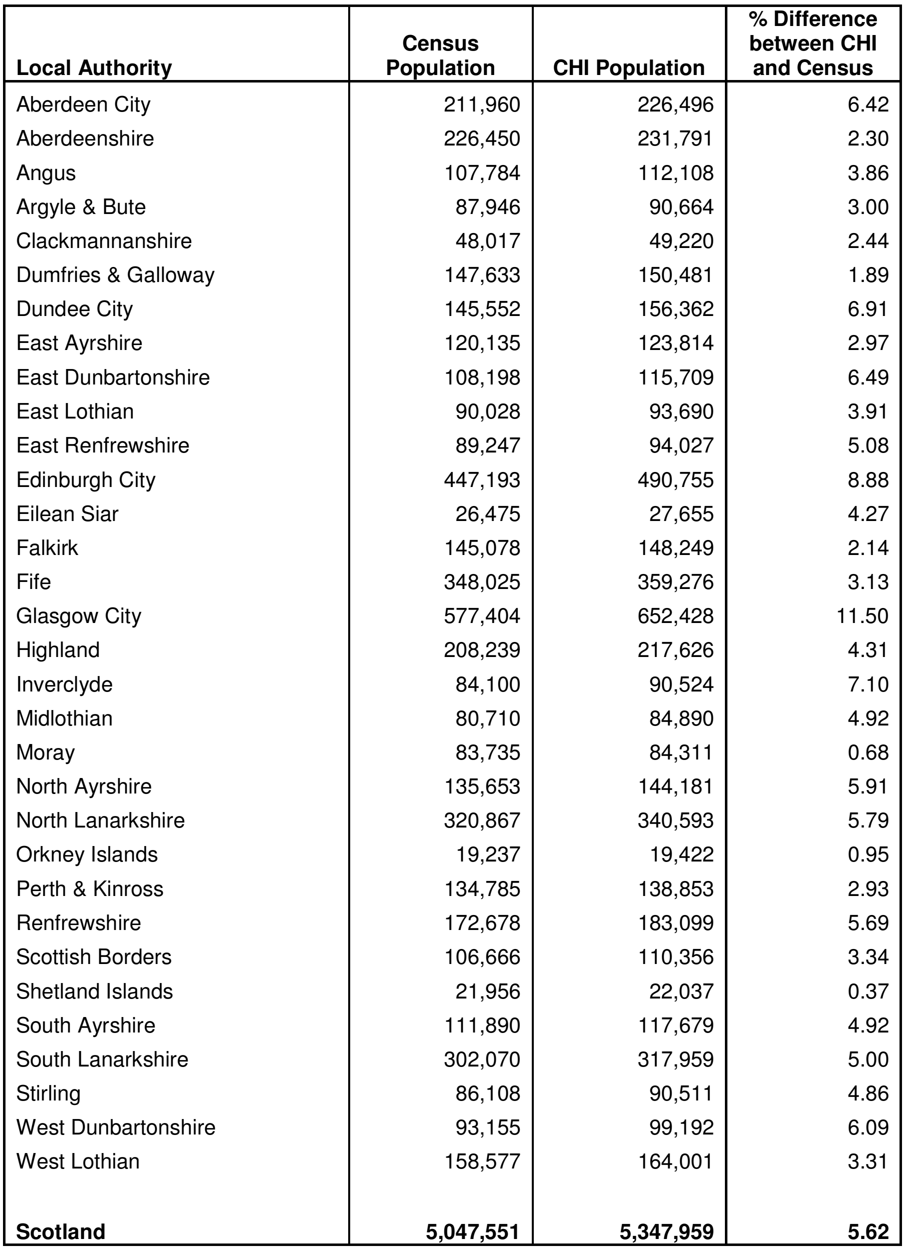
Figure 2.1: The Census and CHI population, by local authority (in April 2001)



A properly maintained administrative system has a quality assurance aspect such that there is a constant monitoring process to ensure the reliability and validity of the data. However, validity and reliability in the CHI present ongoing challenges. Apart from the issues surrounding migration across different health boards, the CHI has been found wanting in a number of situations. In 1996 the CHI had an overall inflation of about 8%, mainly attributable to people who had moved or died (Womersley (1996)). In fact, at a small area level these inflation rates were revealed to be as high as 20%. There were also age differentials in the inflation rates, with over 75s and 20-30 year olds having the highest inflation rates. This does make sense, as people in their 20s (especially males) are more likely to be transient. Also, it was found that married women are duplicated when they change their names. A few years later, after the 2001 One Number Census was completed a similar exercise was undertaken by the General Register Office for Scotland (GROS) to compare the CHI to the Census results and they estimated a discrepancy of 5% between the mid-year estimate of the population and the CHI counts of patients registered (see Table 2.4).

The GROS results (presented in Figure 2.1 and Table 2.4) do provide some support to Womersley's results that at higher geographical levels the CHI population counts are fairly similar to the census population estimates, although they show that in all local authorities the CHI population estimate is always higher than the Census, and more so in cities. So in effect the CHI does count young people, but unfortunately this occurs in the wrong place. Womersley estimated his overcount figures by cross-checking whether people on the CHI were resident at the address given on their records. On the basis of this there is some potential of using information on the CHI to provide data about particularly hard to enumerate people.

Table 2.4: Difference between the Census and CHI population, by local authority (in April 2001)



Local Authority	Census Population	CHI Population	% Difference between CHI and Census
Aberdeen City	211,960	226,496	6.42
Aberdeenshire	226,450	231,791	2.30
Angus	107,784	112,108	3.86
Argyle & Bute	87,946	90,664	3.00
Clackmannanshire	48,017	49,220	2.44
Dumfries & Galloway	147,633	150,481	1.89
Dundee City	145,552	156,362	6.91
East Ayrshire	120,135	123,814	2.97
East Dunbartonshire	108,198	115,709	6.49
East Lothian	90,028	93,690	3.91
East Renfrewshire	89,247	94,027	5.08
Edinburgh City	447,193	490,755	8.88
Eilean Siar	26,475	27,655	4.27
Falkirk	145,078	148,249	2.14
Fife	348,025	359,276	3.13
Glasgow City	577,404	652,428	11.50
Highland	208,239	217,626	4.31
Inverclyde	84,100	90,524	7.10
Midlothian	80,710	84,890	4.92
Moray	83,735	84,311	0.68
North Ayrshire	135,653	144,181	5.91
North Lanarkshire	320,867	340,593	5.79
Orkney Islands	19,237	19,422	0.95
Perth & Kinross	134,785	138,853	2.93
Renfrewshire	172,678	183,099	5.69
Scottish Borders	106,666	110,356	3.34
Shetland Islands	21,956	22,037	0.37
South Ayrshire	111,890	117,679	4.92
South Lanarkshire	302,070	317,959	5.00
Stirling	86,108	90,511	4.86
West Dunbartonshire	93,155	99,192	6.09
West Lothian	158,577	164,001	3.31
Scotland	**5,047,551**	**5,347,959**	**5.62**

Source: Ganka Mueller, Demography Division, General Register Office for Scotland

Finally, in theory the CHI should register any time an update to patients' records is made, and so for transient young males it may be possible to know their last address - this information would prove very valuable when matching to the Census and Survey. However, at lower levels of geography, for example at the postcode level, the CHI will not be very accurate. Nonetheless, the CHI does contain a considerable amount of information (as shown by Table 2.3) and though the CHI poses problems when it comes to deducing from it a comprehensive count of the population at very low geographies due to the levels of erroneous enumerations, it can be useful during the census. In paraphrasing Womersley's concluding remarks;

> *it does seem a shame not exploit this invaluable data source to the full.*

2.5 Measurement of Overenumeration in Censuses

Although not explicitly stated it is assumed that the observed counts in the 2x2 contingency table (see Table 2.1) have been cleaned of any overenumeration. In fact, it may be construed from the assumption that individuals can be matched from capture to capture that every census enumeration is correct. However, in actuality, this may not entirely be true; some enumerations will be erroneous. To that end, most census taking countries undertake some procedures within the census assessment to adjust for these errors. In the UK, for instance, there are a number of clerical matching and data processing techniques that make sure that the counts of those enumerated in either the Census or Survey are devoid of any erroneous counts. Further, the Survey collects data on possible locations of where individuals could have been counted in the Census, and ad hoc adjustments can be made to age-sex dual system estimates (Brown, 2000, page 112). However, in the US, overenumeration is a more serious issue owing to the broad range of erroneous enumerations that exist (e.g. duplicates, fictitious census returns and incorrect census imputations). Accordingly, the US Census Bureau has a part of their accuracy and coverage evaluation that explicitly tackles the estimation of overenumeration. This is done through two surveys, namely the E-sample and P-sample.

The E-sample can be thought of as the initial enumeration of the population (i.e. the 'Census' in the context of the UK) while the P-sample is the subsequent independent survey of the population, and is equivalent to the UK's Census Coverage Survey. In order to obtain an estimate of overenumeration, a sample of census returns are re-visited to check for fictitious data and erroneous or incorrect enumerations. The sample is restricted to those people who fail to match the P-sample records. It is now possible, in theory, to determine for each individual whether they were enumerated in the Census despite being missed in the P-sample, or whether they were erroneously enumerated.

The E-sample and P-sample are used to estimate an adjustment for both the underenumeration and overenumeration by calculating the proportion of the population correctly included in the Census, which is estimated by the P-sample match rate, and the proportion of the Census counts that were correctly included, which is estimated by the E-sample correct enumeration rate (Citro et al., 2004, page 160). The dual system estimate of the population is

$$\hat{N} = (C_{el}) \left(\frac{E_{cor}}{N_e} \right) \left(\frac{N_p}{P_{mat}} \right) \tag{2.1}$$

where C_{el} is the number of census counts that are eligible to be matched to the P-sample, E_{cor} is the number of E-sample persons who were correctly enumerated in the Census, N_e is the total number of E-sample persons, P_{mat} is the number of P-sample persons who match with the E-sample, and N_p is the number of P-sample persons. The equation (2.1) follows from the independence assumption that the probability of being enumerated in the Census is not related to the probability of being enumerated in the P-sample. In other words as a result of the independence between the P-sample and Census, the estimated proportion of P-sample people who match to the Census, $\frac{P_{mat}}{N_p}$, is a good estimate of the proportion of people who were correctly enumerated in the Census, $\frac{E_{cor}}{\hat{N}}$. So

$$\frac{P_{mat}}{N_p} = A_{fac}\frac{E_{cor}}{\hat{N}}$$

where $A_{fac} = \frac{C_{el}}{N_e}$ is an adjustment factor to account for the fact that only a proportion of those counted in the Census are real people, but the matching of the P-sample is across the whole Census[10]. This can be contrasted to Canada where no dual system estimation is employed, nonetheless the final population estimates are adjusted for both underenumeration and overenumeration through a Reverse Record Check. The assumption is that, after matching, the number of persons in neither the Census nor the Reverse Record Check is expected to be negligible relative to N (Martel and Caron-Malenfant (2007)), so the estimate of the population is

$$\hat{N} = N_C + \tilde{U} - \tilde{O}, \tag{2.2}$$

where N_C is the number counted in the Census, $\tilde{U}$ is the estimate of undercount and $\tilde{O}$ is the estimate of overcount.

The US has conducted extensive research on alternative methods of coverage assessment. One such method, referred to as CensusPlus (Mulry and Griffiths (1996)) works by undertaking an intensive enumeration of a sample of the population. But whereas in dual system estimation the results from this sample are matched to information from the census, CensusPlus seeks to obtain as accurate a count of the sample population as possible through using the best staff and resources available. A crucial assumption here is that complete coverage of the sample can be achieved. However, Judson (2006) suggests that, although appealing, complete coverage in the sample is too heroic (and possibly untenable) an assumption to make.

These techniques detailed above are very different to the one being proposed in the thesis. The proposal here is to use contingency table analysis to fit different models in order to directly obtain estimates of population size that are adjusted for the existing overenumeration and underenumeration.

2.6 Contingency Table Analysis

The contingency table framework of the capture-recapture information has led to the application of more sophisticated statistical theory and inferential procedures. The work by Darroch (1958) and Darroch (1962) laid the foundations of the mathematical framework of this topic, and was largely reliant on the seminal Bartlett (1935) paper. His eponymous result, the Bartlett criterion, allowed the investigation of the associations in complex contingency tables, starting with a simplistic 2x2 (first-order) case (as shown in Table 2.5)

[10] cf: Under DSE, $\frac{\text{count in first sample}}{\text{population total}} = \frac{\text{count in both samples}}{\text{count in second sample}} \Rightarrow \frac{n_{1+}}{\hat{N}} = \frac{n_{11}}{n_{+1}}$

and builds up to a more general case; specifically, the Bartlett criterion extends tests for independence across multi-dimensional contingency tables.

Table 2.5: Two sample capture-recapture

		Second Sample	
		Counted	Missed
Second	Counted	n_{11}	n_{10}
	Missed	n_{01}	n_{00}

Prior to this paper the existing tests of independence were based on Yule and Pearson's work which only strictly applied to 2x2 contingency tables (see Yule (1900) and Pearson (1900)). Further, it was here that a methodology was put in place for the computation of the maximum likelihood estimates for contingency tables.

The association in the table can be measured by the sample odds ratio $\hat{\theta}$, given by

$$\hat{\theta} = \frac{n_{00}n_{11}}{n_{01}n_{10}}. \tag{2.3}$$

When the population odds ratio $\theta = 1$, the first and second samples are independent and this results in the dual system estimator of the missing cell,

$$\hat{n}_{00} = \frac{n_{01}n_{10}}{n_{11}}. \tag{2.4}$$

For the 2x2x2 case, as shown in Table 2.6, the ratio of the odds ratios in the layers of the contingency table can be used to measure the association.

Table 2.6: Three sample capture-recapture

		Third Sample			
		Counted		Missed	
		Second Sample		Second Sample	
		Counted	Missed	Counted	Missed
First Sample	Counted	n_{111}	n_{101}	n_{110}	n_{100}
	Missed	n_{011}	n_{001}	n_{010}	n_{000}

Simply put, assuming that there is no three-factor interaction effect (i.e. all pairs of samples may exhibit dependence but the amount of dependence in each pair is assumed to remain unaffected when conditioned on the third sample), then Bartlett's criterion shows that the test for three-factor interaction effects is equivalent to

$$\pi_{000}\pi_{011}\pi_{101}\pi_{110} = \pi_{001}\pi_{010}\pi_{100}\pi_{111} \Rightarrow \pi_{000} = \frac{\pi_{001}\pi_{010}\pi_{100}\pi_{111}}{\pi_{011}\pi_{101}\pi_{110}}.$$

Forcing this relationship on the cell counts gives

$$n_{000}n_{011}n_{101}n_{110} = n_{001}n_{010}n_{100}n_{111}. \tag{2.5}$$

The missing cell can therefore be written as

$$\hat{n}_{000} = \frac{n_{001}n_{010}n_{100}n_{111}}{n_{011}n_{101}n_{110}}. \tag{2.6}$$

As in the independence assumption in the 2x2 case, it is required to make the untestable assumption that there is no three-factor interaction here[11]. Another non-trivial point to make from equation (2.5) is that it is not possible for $\hat{n}_{000} = 0$; this implies that there is always someone in the missing cell. Put differently, no combination of the three lists has full coverage of the population.

For categorical data presented in a contingency table, the log-linear model is a good tool in the analysis of the relationships of the variables. It is especially useful in testing the hypothesis that the cell counts in the cross-classified table are consistent with statistical independence. Regression models can be used, but the advantage a log-linear model has is that it can be employed to address the different types of departures from statistical independence. The log-linear model is relatively new, compared to regression models but according to Sobel (1995) its origins can be traced to the work of Pearson (1900) and Yule (1900). Pearson's χ^2 test for independence and Yule's $\mathbf{Q}$ measure of association are both similar to the odds ratio that underlies log-linear models. Birch (1963) was the first to express the log-linear model in its current form, and developed the basic asymptotic theory under Poisson and multinomial sampling.

In a census application Fienberg (1972), using Darroch (1958) as a basis, fitted a log-linear model to the incomplete contingency table that results from a capture-recapture experiment. He derived different models to fit the observed data under different dependency assumptions. An estimate of the missing cell is then found under the simplest (parsimonious) model for the data, and this consequently yields an estimate of the total population size. Chapter 6 of Bishop, Fienberg and Holland (1975) gives a comprehensive overview of capture-recapture population estimation. The book also gives the asymptotic variance results for the population estimate under different log-linear models. These variance estimators are based on the assumption of normality of the population size estimate derived in Sanathanan (1972a) and Sanathanan (1972b) using maximum likelihood estimation. However, maximum likelihood estimation has a disadvantage in its assumption of asymptotic normality in capture-recapture. Seber (1982) and Agresti (1994) showed that the distribution of the population size estimator can be markedly skewed, and so the assumption of asymptotic normality may be flawed. Buckland and Garthwaite (1991) suggested a boostrapping procedure as a method of quantifying the precision of the population size estimator. Cormack (1992), Agresti (1994) and Coull and Agresti (1999), on the other hand suggest a profile likelihood function that views the maximized likelihood as a function of the unobserved cell count. In fact, Coull and Agresti (1999) found that rather than being centred in the interval, it is common for the population estimator to be nearer to the lower end of the interval which is indicative of the skewness. Further, these non-normal confidence intervals have a feature that they are bounded below by what is observed.

[11] Brown, Biemer and Judson (2006) suggest moving to a 'quadruple system' model.

2.7 Latent Class Models

An alternative approach to estimating the population size assumes that individuals cluster into latent classes, such that individuals within the same class have the same catchability. This is not a new concept, as this is what was suggested by Chandrasekar and Deming (1949) when they spoke about post-stratification. They proposed that dividing the population into groups based on age, sex and geographical region, results in subgroups within which the individuals roughly behave in the same manner with regards to their capture probabilities. This is the basis of their homogeneity assumption.

However, there are some cases where even after stratification by observable traits, some residual heterogeneity continues to exist in the stratum. The methods mentioned above do seek to model this extra dependence, but the latent class approach is a different way of achieving the same objective - in fact the Rasch model can be thought of as being a latent class model with potentially as many latent classes as there are individuals. Thus, when it is not possible to account for the heterogeneity of capture probabilities by taking account of the observed covariate information, a latent class model can be found that classifies the individuals into a small number of groups with homogeneous capture probabilities and list independence, conditional on the latent classes. So for example in a case where there are two latent classes in an animal capture-recapture experiment, the population can be treated as a mixture of two types - ones that show an aversion to trapping (hard to count) and the others who show an attraction to trapping (easy to count). Within each latent class it is assumed that the animal captures are independent.

From another view point, the underlying assumption of independence implies that the probability of enumeration in the Survey, given enumeration in the Census is identical to the probability of enumeration in the Survey given that the individual was missed in the Census. This assumption can be too strong and fail to hold, because of the heterogeneity in individuals' probability of being captured in either of the two processes, as mentioned in the earlier discussion of correlation bias. The normal way of trying to rectify this situation is by post-stratification, also mentioned earlier. However, there are some cases where it may be inevitable that the post-stratification fails to account for all the heterogeneity, and thus the Census and Survey inclusion probabilities vary from person to person within a post-stratum. The latent class approach is therefore an appropriate way of handling this heterogeneity. Although the data may have been post-stratified using the demographic, socio-economic and household factors known to affect underenumeration[12], some individ-

[12] The Hard-to-Count Index used in the 2001 One Number Census utilised the variables known to be associated with census underenumeration, such as high levels of multi-occupancy and private rented accommodation, but not necessarily those factors that increase the enumerator work load such as a large geographical area. Nevertheless, inasmuch as it was to sought to include all possible variables, some factors were missed that were later found to have an impact on the the level of underenumeration, for example the number of second homes and the rate of redevelopment and regeneration, as evidenced in parts of Westminster and Manchester (see Office for National Statistics (2004)).

uals with differing levels of 'catchability' could be placed within the same post-stratum. The interpretation is that there is some underlying, hidden variable which when taken into account successfully splits the population into homogeneous sub-groups.

Latent class analysis is justified in epidemiological capture-recapture studies because in most cases the lists used in the measurement of the population are set up for various purposes, which could introduce relationship structures that cannot always be modelled using the observed covariates. There is also unobserved heterogeneity due to complex patterns that may exist as a result of the phenomenon under study. In the context of a three-system capture with the Census, Survey and Administrative List, the data collection methodologies of each of the systems can be envisaged as different. Although the Census and Survey processes could be construed to be similar, the Administrative List does have some peculiarities that are only applicable to it.

In the 2001 One Number Census it was assumed that there were no erroneous enumerations in the Census and Survey counts. This was done by clerical matching to ensure that any duplicates were duly corrected for. In actuality, albeit the matching process managed to remove the obvious duplicates and errors (for example census forms returned for Donald Duck at Disneyland), it was much more difficult to determine where to correctly place some other people.

Two particular sub-groups were identified. The first group were children of divorced parents, who split their time between both parents. In some cases, the matching process using the relationship matrix[13] did manage to identify these children. However, the difficulty arose as to where to place these children. The information from the relationship matrix became difficult to match on when the parents had remarried. There were other complex relationship patterns that the matrix could not cope with, especially when the households were large.

The second group were students. Following the consultation process after the 1991 census it was decided to capture information about both term-time and home addresses of students in 2001. The matching process worked on the basis that the student's parents had put them on the census form. There were some obvious difficulties - students fall within the most difficult to enumerate age group, so the census returns from their parents had them enumerated, but in most instances at the wrong location. For areas where students form a large proportion of the population the potential impact of failing to count them in the correct location poses a lot of difficulty when it comes to resource planning.

What this shows is that when a third list is introduced into the population estimation process the plausibility of erroneous enumerations becomes a very real issue. From the review of literature, the most probable administrative register that can be used as the third

[13] In the 2001 Census form the relationship matrix showed how household members were related to each other. However, it must be noted here that the recent rounds of census tests (carried out in 2006 in Scotland and 2007 in England and Wales) did not have a relationship matrix on the census forms.

list is the National Health Service Central Register. However, it was not set up originally to serve as a population register. During the set-up of the National Health Service, a patient register was established for the primary purpose of organising remunerations to General Practitioners; so the more patients the practice had, the more money it received from the government. The onus was on the practice to inform the NHS that a person had died or moved. With increasing GP patient lists this house-keeping process became less frequent, as other duties took priority.

When a general audit was carried out in the 1970s (after the NHSCR began to be touted as a population register) it was found that a large number of dead people were still on the register (Redfern (1989)). The checks used were quite crude - only persons over the average life expectancy were followed up, so it could be construed from this that the figures may actually under-estimate the extent of erroneous enumerations. There is also the problem of relying on people remembering to register with a new medical practice as soon as they move. However, with the hecticness surrounding house moves this features probably at the bottom of the list. People will only remember to register when they are ill - this poses a problem for young adult males who are known to visit their GPs less often.

The basic premise of latent class analysis as applied in a triple system setting is that the observed covariation between the Census, Survey and Third List is actually better represented by each of the Census, Survey and Third List's relationship with an unobserved latent variable. In fact, (McCutcheon, 1987, page 11) argued that one way of suggesting that a latent model might be appropriate could be when the variables are so inter-related to suggest that any observed patterns of associations could not be wholly attributed to chance only. In other words, for the triple system application, even with n_{000} observed, the best fitting model will be the one that has all the interaction terms. So it would appear that the Census, Survey and Third List are correlated. However, bringing in a latent variable results in there being independence between the Census, Survey and Third List (after controlling for the latent variable). In consequence, the latent variable is the 'true' source of the observed associations between the Census, Survey and Third List. Since this latent variable is unobserved, the crucial part of the analysis lies in the interpretation of the classes of the latent variable. In triple system estimation, the latent variable is meant to characterize unobserved (unexplained) heterogeneity, and this could be due either to a failure of the post-stratification mechanism or capture error.

The two interpretations of the latent class (i.e. enumeration difficulty or enumeration error) are very different and lead to conflicting population estimates. In the first case, the latent class represents people according to how difficult they are to enumerate. Therefore, everyone is included in the total population. In the second case, it represents whether or not the enumerations are real or erroneous people. Here the total population is only those deemed to be real; any erroneous counts can be then removed. The decision as to how to interpret the latent classes after analysis is therefore very subjective, and is one of the major drawbacks to the application of latent class modelling in census estimation.

Latent variable modelling is often viewed as a dubious exercise fraught with unverifiable assumptions and naïve inferences (Skrondal and Rabe-Hesketh, 2004, page 6). Nevertheless, considering latent variables in the analysis is useful in providing explanations as to how the different systems used in census estimation work. For example, a latent class analysis on a cross-section of the data may subdivide the data into contiguous groups, separating out the 'different types' of people. Looking at the make-up of these groups (using the observed covariates such as demographic and household information) it may be possible to give simple interpretations to these groups, which in turn can provide valuable insight to the whole population, or provide better post-stratification.

There are two different parameterizations of the basic latent class model - the first attributed to Goodman (1974) and the second to Haberman (1979). The Goodman parameterization is based on conditional probabilities, while the Haberman parameterization uses log-linear models. A latent class model can, in essence, be regarded as a log-linear model in which only marginal totals are observable. Since the latent variable is unobserved, the counts within the latent classes are unknown - i.e. only the marginal totals summed over the latent variable are known. In a capture-recapture application the marginal total over the unknown (missing) cell is additionally unobserved.

In either parameterization, there are two key assumptions used to specify the latent class model. Firstly, the population is assumed to consist of a set of mutually exclusive and exhaustive homogeneous subgroups (in the latent model being considered during the course of the thesis there are two classes). These groups are the latent classes, by definition. Secondly, within a given latent subgroups all observable indicators are statistically independent. This is what is known as local dependence - in essence the observable variables are conditionally independent given the categories of the latent (unobserved) variable. More recently Skrondal and Rabe-Hesketh (2004) classify latent class analysis into exploratory or confirmatory. Exploratory latent class analysis makes no a priori restrictions on the parameters of the model, while confirmatory analysis does place some restrictions. In confirmatory analysis prior information - from substantive theory or previous results - is used to determine the latent classes. They define exploratory analysis as an inductivist method to discover the optimal set of latent classes. Therefore, whether an exploratory or confirmatory approach is taken will influence how the latent classes are defined. Notwithstanding these concerns, it can be assumed that for a suitably post-stratified population, the latent class model is an intuitive solution to coping with both dependence and overenumeration.

2.8 Bayesian Methods

Bayesian methods for capture-recapture have been scarce, and this is surprising given that the population estimate is found by updating information from previous capture histories. An individual's previous capture information plays a pivotal role in the assumptions in determining the estimate of the population size. This is the main ethos of Bayesian statistics as it provides a mathematical framework for revising knowledge based on prior information. Further, as mentioned earlier, the classical approaches have encountered problems when it comes to the formulation of confidence regions for the population estimator. The classical method for obtaining a confidence interval entails an approximation and an assumption. Initially, an approximation for the standard error of an approximately unbiased estimate is obtained. Then the point estimate is assumed to have a normal distribution and symmetric confidence intervals are constructed. Even under profile likelihood estimation, a distributional assumption is required, and hence the intervals are not exact. In the Bayesian paradigm, however, given the observed data likelihood and the prior, the intervals are exact. On the other hand, frequentists would argue that Bayesian intervals are also inexact since they rely on the specification of the prior. What is apparent is that with the advancement of computationally intensive statistical methods, Bayesian inference has an important role in capture-recapture population size estimation (King and Brooks (2001)).

Castledine (1981), Gazey and Staley (1986), Smith (1988), Smith (1991) and Zelterman (1988) are the earliest Bayesian analyses of capture-recapture. In a census context, Zaslavsky (1993) used a Bayes-type approach (i.e. loss functions) to produce estimates of the population using data from the 1990 Census, Dual System and Evaluation Study. More recently Nandram and Zelterman (2007), using an incomplete 2x2x2 contingency table of Spina Bifida cases in New York from 1969-1974, utilized rejection-sampling and a Bayesian log-linear model to estimate the population size. The aim was to investigate the prevalence of Spina Bifida based on data from birth, death and medical rehabilitation records. In classical model selection, there is the risk that the model chosen may be wrong, but there is no way of determining the validity of the results. So the advantage of the Bayesian paradigm within the capture-recapture situation is that each model can be given a posterior probability which is representative of how likely the model is, in turn allowing model uncertainty to be (explicitly) incorporated into any decisions or predictions.

Thus to estimate the population size using a log-linear model, the best-fitting model is selected and the missing cell count that maximizes the likelihood under the chosen model is estimated. There are three problems with this approach. In the first instance, as the number of sources grows, the number of possible models increases exponentionally. It can therefore become difficult to differentiate between models. Secondly, the maximum likelihood estimator for the missing cell, and relatedly the estimate of the total population size is often sensitive to the choice of model. It can become difficult to assign much confidence to the results obtained. Accordingly, the Bayesian paradigm seeks to overcome these problems by calculating the posterior probability of each model. Thirdly, there are some times when more than one log-linear model can be found that fits the data well. The classical population size estimation methods do not take advantage of information on the size of the population, which may be available a priori, and cannot incorporate covariate information easily.

Madigan and York (1997) use a Monte Carlo Markov Chain (MCMC) approach for investigating model uncertainty, while Dellaportas and Forster (1999) use a reversible jump MCMC approach. In the first approach only decomposable models are given a posterior probability, so not all models are considered. For models with a large number of variables, the total number possible models will make the calculation of model probabilities prohibitive. Thus a reversible jump MCMC is a quicker way of considering all the models while speeding up the computational process. King and Brooks (2001) used this approach in a capture-recapture application.

The most current use of Bayesian inference in regards to human population censuses is the work for the US Census Bureau by Stuart and Zaslavsky (2002) and Stuart and Judson (2003). In a triple system census - with data from the Census, Post-Enumeration Survey and Administrative Records - they relax the assumption that the population is closed and propose a hierarchical model, seeking to tackle the problem of measuring transiency. The proposed model has three levels, with the first level describing the probabilities of observation in each of the available systems, the second level describing the migration process, and finally the third level describes the priors governing the global parameters. Their motivation was the fact that one of the drawbacks of using administrative records is that their coverage periods may not often coincide with the date of the census. The hierarchical model developed looked to model migration by predicting whether someone is still a resident on census day given that they appear on more than one recording system. The advantages of this are plentiful. First, if the administrative records are available nationally, the model developed can be used to provide small area underenumeration estimates across the whole country. These estimates are based on the whole population and not on the sampled blocks and therefore the final population estimates rely less on synthetic estimation. Second, it may be used to add or subtract people for whom there is evidence that they were or were not in the area on census day. Theoretically, the post-enumeration survey may be used to accomplish this task but, again, the estimates will only apply to sampled blocks and not the entire population. Additionally, there are fewer assumptions of homogeneity across areas, and so the local underenumeration estimates can be more reliably obtained.

There are, however, some limitations, the first being that the independence assumption across systems is unrealistic, and dependence cannot easily be implemented in the model. Secondly, there was the susceptibility of the model to over-estimate the population size, when file coverage and migration parameters were wrongly estimated - i.e. the implications of model mis-specification had not been accounted for. Lastly, the assumption of high quality perfect matching and non-duplication is ambitious. In reality matches are imperfect and dates and address information may be wrong. But, more positively, the hierarchical model can be useful in targeting individuals for more intensive follow-up. Individuals with high or low probabilities of census day residency could be given comparative follow-up probabilities. The advantage of such an approach is that resources could be targeted to individuals that have ambiguous results. The current extension of the hierarchical model considered by the US Census Bureau looks at developing the model at a household level, with the hope that it would lead to better estimation - most moves are, after all, at a household level (Stuart and Judson (2003)). Also, migration is often dependent on geographical area, thus incorporating local information and migration patterns could be beneficial. Further extensions of the general hierarchical model could assimilate inexact matching, erroneous enumerations, heterogeneity of capture probabilities and heterogeneous forms of migration.

2.9 Demographic Analysis as an Alternative Method of Coverage Evaluation in Censuses

Census methodologies can be broadly divided into

(a) Traditional Enumeration: - a canvass of all individuals in the population is undertaken.

(b) Register-based Enumeration: - information about individuals is combined from a number of different administrative sources.

(c) Survey-based Enumeration: - population estimates are derived from nationally representative survey data.

Each methodology has its advantages and disadvantages. Nevertheless, in general, the choice of which methodology a country uses to produce population estimates is dependent on national circumstance and resources. More often than not, countries use a mixture of methodologies in order to produce population estimates that are considered accurate and reliable. The UK, for example, undertake a traditional enumeration but in addition to using a post-enumeration survey for coverage assessment also make use of aggregate-level administrative source data to quality assure the census results.

In the preceding discussion, there has been a focus on dual and triple system estimation as methods for assessing the coverage of the population census. It has to be mentioned here that there other non-statistical methods available, the most widely used being *demographic analysis*. In demographic analysis the estimate of the population is arrived at by rolling forward the most recent population estimate allowing for births, deaths and net migration. This can be given as

$$P_{t+1} = P_t + B_t - D_t + M_t^I - M_t^E,$$

where P_{t+1} is the new population estimate, P_t the current population, B_t the number of births, D_t, the number of deaths, M_t^I the number of immigrants and M_t^E the number of emigrants. In reality, although the number of births and deaths can be estimated to a fairly decent degree of accuracy from administrative resources, the estimation of net migration is far from easy.

In the UK (documented) migration data is primarily obtained from two sources - the UK Borders Agency (UKBA) and the International Passenger Survey (IPS). The UKBA is the government department that is responsible for managing migration, while the IPS is a randomized face-to-face survey of individuals entering or leaving the UK by air, sea or the Channel Tunnel. It conducts roughly a quarter of a million interviews of passengers throughout the year; this equates to approximately 1 in 500 passengers (Office for National Statistics (2007)). Further, the IPS was designed for the compilation of balance of payments and to provide information about tourism, as well as obtain the characteristics and numbers of migration in and out of the UK. Therefore, it is clear that the IPS cannot provide sufficiently detailed micro-level migration information, particularly at lower geographical levels due to the design and small sample size.

At a national level (or macro-level), however, demographic analysis can be useful. In both the US and UK demographic analysis is used to check the assumption of independence between the initial census enumeration and post-enumeration survey (see Wolter (1990), Bell (1993) and Brown, Abbott and Diamond (2006)). The estimate of the national level population post-stratified by age and sex obtained from the Census and Survey can be compared to the historical data on births, deaths and migration. Another important aspect of demographic analysis is the ratio of males to females in the population, otherwise known as the *sex ratio*. The theory is that the sex ratios obtained under the census should be roughly similar to those under demographic analysis. If there is a discrepancy, then the belief is that this discrepancy is evidence of dependence between the Census and Survey. Therefore, dependency adjustments can be applied to national age-sex population estimates. This is what happened in the 2001 UK Census.

Demographic analysis can be - and is - used as a method of census coverage evaluation. Nationally, it can be used to quality assure the census figures, in particular when there is reason to doubt the independence assumption. Subnationally, however, demographic analysis encounters problems due to lack of reliable individual level demographic data.

2.10 Conclusion

In the 2001 UK One Number Census, dual system estimation was used to estimate the total population size including those missing. The two systems considered were the Census and the Census Coverage Survey, and relied on the two basic assumptions of homogeneity and independence. Another assumption was that there was relatively high coverage across the population achieved. A failure of any of the initial assumptions, combined with low levels of coverage introduces bias into the population estimates. The factors that made enumeration difficult in 2001 - changing demographic, socio-economic, complex household structures and public attitudes - are expected to feature more strongly in 2011. Thus data from a third source - an administrative list - has been proposed as a means of correcting for this bias. This becomes triple system estimation where individuals are cross-classified according to their presence or absence in each of three lists: the Census, Survey and the List, and has far less restrictive assumptions. Administrative lists are unique in terms of their accessibility, inclusiveness and flexibility and so can reasonably be considered as being independent of the Census and Survey.

The most feasible administrative list for triple system estimation, in the format envisaged by this thesis, is the National Health Service patient register. However, the current UK-wide health registry data is not very useable because of the lack of unique person identifiers. Nonetheless, Scotland's Community Heath Index (CHI) has demonstrably shown that, owing to a concerted effort in the 1970s to ensure that every person living in Scotland (or for that matter, who has lived in Scotland for an extended period) has a unique health number, there is some promise. Even accounting for the fact that young males, who have been found to be most problematic to count in previous censuses, are less likely to visit their GPs, there is a wealth of information that can be harvested from the CHI. However, admittedly, the introduction of the third list brings with it some added complications. Overenumeration, assumed to be negligible in previous censuses, is introduced into the cell counts due to imperfections in the CHI. Nevertheless, the idea is that a log-linear model (either in a Classical or Bayesian paradigm) can be used to estimate the population size after matching information gathered from the Census, Survey and CHI[14].

There is another approach, using latent class models, to estimate the population size. The latent class approach becomes particularly useful when there is unobserved heterogeneity. This unobserved heterogeneity can be due to a failure of the post-stratification mechanism or there being some erroneous enumerations. The latent class model specified under the Haberman parameterization can be written as a log-linear model.

[14] The chosen blocks are presumed to have reasonably high coverage for both the Survey and the CHI. Matching over these blocks is less prohibitive than carrying out the process over Scotland, or the UK for that matter. On another note, the time and difficulty it takes for this to be performed for the sampled blocks will give some indication as to how long it will take for the whole of the UK.

Chapter 3

Population Estimation in Capture-Recapture Models

3.1 Introduction

The aim of this chapter is to unify the methodology of capture-recapture from a number of sources, namely Cormack (1972), Cormack (1989), Fienberg (1972), Fienberg (1992), Goodman (1974), Smith (1988), El-Khorazaty et al. (1977), International Working Group for Disease Monitoring and Forecasting (1995a), International Working Group for Disease Monitoring and Forecasting (1995b), Biemer et al. (2001a), Biemer et al. (2001b), Lazarsfeld and Henry (1968), Chapter 6 of Bishop, Fienberg and Holland (1975), Chapter 10 of Haberman (1979), Seber (1982), McCutcheon (1987), Hagenaars (1993), Chapter 9 of Little and Rubin (2002), Chapter 8 of Agresti (2002), Chapter 12 of Congdon (2005) and Brown (2000).

Some of the results have been presented in the above texts but what is different here is that they have been brought together under the same framework and applied to census estimation, with the hope of giving some background to the techniques used in the thesis. Furthermore, although the chapter does review the current capture-recapture literature, there are some new ideas presented here, particularly those pertaining to extending the latent class methodology to coping with both dependency and overenumeration.

Throughout the thesis unless otherwise stated, an assumption is made that the population has been suitably stratified so that individuals have been divided into distinct and non-overlapping sub-populations. This ensures that with respect to individuals in different sub-populations there is heterogeneity, but within each sub-population there is internal homogeneity - meaning that the individuals in the sub-populations have the same inclusion probabilities. This is fairly important as the event of an individual's inclusion or exclusion from one system may be different for different types of individuals. However, after the stratification there is homogeneity across individuals. Moreover, this assumption implies that the bias in the population estimates can be fully attributed to list dependence.

3.2 Dual System Estimation

In order to introduce the notation, it is best to start with the simple two-sample capture recapture problem. Let N be the total number of individuals in the population, n_{1+} the number of individuals in the first sample and n_{+1} the number in the second sample. Similarly, n_{11} is the number of individuals observed in both samples, n_{10} is the number of individuals observed in only the first sample, and n_{01} is the number of individuals observed in the second sample but not the first. The data can be arranged in the form of a 2x2 contingency table, as shown in Table 3.1 where n_{00} is the count corresponding to the missing, and unobservable, cell.

Table 3.1: Two sample general capture-recapture problem

		Second Sample	
		Counted	Missed
First Sample	Counted	n_{11}	n_{10}
	Missed	n_{01}	n_{00}

Denote the number of individuals observed in the two samples by n. Thus $n = n_{11} + n_{10} + n_{01}$. The dual system model assumes independence between the two samples, which implies that the probability of being in the $(i,j)^{th}$ cell, π_{ij}, is the product of the marginal probabilities π_{i+} and π_{+j}, where $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$. So in addition to assuming independence across individuals, the dual system model makes an explicit assumption of independence within individuals. Also let π_{11} be the probability of an individual being observed in both the first and second samples, π_{1+} the probability of an individual being in the first sample and π_{+1} the probability of being in the second sample.

Suppose we assume independence between the two samples, then $\pi_{11} = \pi_{1+}\pi_{+1}$ since $\pi_{ij} = \pi_{i+}\pi_{+j}$ under independence. Further, assuming that n, the total number of individuals observed, is fixed, then the cell counts are multinomially distributed with probability function

$$\begin{aligned}&\binom{n}{n_{01}, n_{11}, n_{11}} \frac{(\pi_{11})^{n_{11}} (\pi_{10})^{n_{10}} (\pi_{01})^{n_{01}}}{(\pi_{11} + \pi_{10} + \pi_{01})^{n}} \\ &= \binom{n}{n_{01}, n_{11}, n_{11}} \frac{(\pi_{+1}\pi_{1+})^{n_{11}} \left[\pi_{1+}\left(1 - \pi_{+1}\right)\right]^{n_{10}} \left[\pi_{+1}\left(1 - \pi_{1+}\right)\right]^{n_{01}}}{\left[1 - \left(1 - \pi_{1+}\right)\left(1 - \pi_{+1}\right)\right]^{n}}.\end{aligned}$$

Now, since the probability of being observed in at least one of the two samples is $(1 - (1 - \pi_{+1})(1 - \pi_{1+}))$, another way of re-expressing the cell counts is as a binomial distribution, with probability function

$$\begin{pmatrix} N \\ n \end{pmatrix} \left[1-\left(1-\pi_{+1}\right)\left(1-\pi_{1+}\right)\right]^{n}\left[\left(1-\pi_{+1}\right)\left(1-\pi_{1+}\right)\right]^{N-n}. \tag{3.1}$$

Obviously, N is unknown but for given values of π_{1+} and π_{+1}, it is possible to find the value of $\breve{N}$ that maximizes equation (3.1) to be

$$\breve{N}=\frac{n}{1-\left(1-\pi_{1+}\right)\left(1-\pi_{+1}\right)}. \tag{3.2}$$

Using the fact that the maximum likelihood estimates of the probabilities π_{1+} and π_{+1} are

$$\hat{\pi}_{1+}=\frac{n_{11}}{n_{+1}}$$

and

$$\hat{\pi}_{+1}=\frac{n_{11}}{n_{1+}},$$

then the maximum likelihood estimate of the population size is

$$\hat{N}=\frac{n_{1+}n_{+1}}{n_{11}}. \tag{3.3}$$

Trivially, $\hat{N}$ and $\breve{N}$ can be shown to be the same, on the proviso that $\pi_{+1}=\hat{\pi}_{+1}$ and $\pi_{1+}=\hat{\pi}_{1+}$. Further, the estimator (3.3) is what is known as the Lincoln-Petersen estimator (see Seber (1982)).

Equations (3.1), (3.2) and (3.3) are the same as the respective equations (6.2-5), (6.2-6) and (6.2-7) in Bishop, Fienberg and Holland (1975).

When two samples are not independent, then $\pi_{11} \neq \pi_{+1}\pi_{1+}$,
and

$$\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}=\gamma. \tag{3.4}$$

This can be re-expressed as

$$\frac{\pi_{11}\left(1-\left(\pi_{1+}+\pi_{+1}-\pi_{11}\right)\right)}{\left(\pi_{1+}-\pi_{11}\right)\left(\pi_{+1}-\pi_{11}\right)}=\gamma \tag{3.5}$$

and yields the quadratic

$$\pi_{11}{}^{2}\left(1-\gamma\right)+\pi_{11}\left(1-\pi_{+1}-\pi_{1+}+\gamma\left(\pi_{+1}+\pi_{1+}\right)\right)-\gamma\pi_{1+}\pi_{+1}=0. \tag{3.6}$$

It is not possible to estimate the dependence unless there is some additional information provided (Bell (1993) and Brown, Abbott and Diamond (2006)), since there are four independent unknowns but three pieces of information available in equation (3.4).

3.3 Triple System Estimation

It is evident that another way of expressing the estimate of the missing cell under dual system estimation is

$$\hat{n}_{00} = \gamma \frac{n_{10} n_{01}}{n_{11}}, \tag{3.7}$$

where γ is the odds ratio or the dependence, which is unknown in dual system estimation. Therefore, it is assumed that $\gamma = 1$ (i.e. the samples are independent of one another).

If there is an additional list, then it becomes possible to investigate whether the assumption of independence in the dual list problem holds. For a three sample capture-recapture, the capture history can be represented in a 2x2x2 contingency table (see Table 3.2), with the missing cell denoted by n_{000}. As in dual system estimation, N and n represent the (unknown) total population size and the (known) number of individuals observed in the samples, respectively.

Table 3.2: Three sample general capture-recapture problem

		Third Sample			
		Counted		Missed	
		Second Sample		Second Sample	
		Counted	Missed	Counted	Missed
First Sample	Counted	n_{111}	n_{101}	n_{110}	n_{100}
	Missed	n_{011}	n_{001}	n_{010}	n_{000}

In a capture-recapture study the sample size is not known or fixed in advance, so for this reason the multinomial distribution is a good working distribution. Furthermore, the multinomial can statistically arise from the product of binomial distributions or of independent Poisson distributions, conditioned on the observed sample size; this result will be useful later on when trying to maximize the multinomial likelihood. Thus, assuming n is fixed, then the seven observed cells $\{n_{111}, n_{110}, n_{101}, n_{100}, n_{011}, n_{010}, n_{001}\}$ follow a multinomial distribution with probability function

$$\binom{n}{n_{111}, n_{110}, n_{101}, n_{100}, n_{011}, n_{010}, n_{001}} \frac{(\pi_{111})^{n_{111}} (\pi_{110})^{n_{110}} (\pi_{101})^{n_{101}} (\pi_{100})^{n_{100}} (\pi_{011})^{n_{011}} (\pi_{010})^{n_{010}} (\pi_{001})^{n_{001}}}{[(\pi_{1++} + \pi_{+1+} + \pi_{++1}) - (\pi_{1+1} + \pi_{11+} + \pi_{+11}) + \pi_{111}]^n}. \tag{3.8}$$

The denominator is the probability of being observed in at least one of three samples, expressed as a function of the observed cells.

Now this distribution can also be expressed as a function involving the unknown parameter, N,

$$\frac{N!}{(N-n)! \prod_S n_{ijk}!} \left(1 - \sum_S \pi_{ijk}\right)^{N-n} \prod_S (\pi_{ijk})^{n_{ijk}} \tag{3.9}$$

where S represents the set of all cells apart from the $(0,0,0)$ cell. The maximum likelihood estimation of the probabilities based on the multinomial distribution specified by equation (3.9) is complicated. Therefore, Sanathanan (1972a) suggests using a conditional maximum likelihood estimation, where the likelihood function can be 'broken' into two terms and then maximized; with the first being conditioned on the second. In essence the parameters are estimated by maximizing the conditional likelihood of the observable capture histories given that the individuals were captured at least once.

In capture-recapture the $\{\pi_{ijk}\}$ are cell probabilities of a multinomial random variable, the π_{000}-cell is unknown and subsequently the population size N is also unknown, but the population size of the observed cells is known to be n. Therefore, the missing count for the π_{000}-cell is $N-n$, allowing the multinomial likelihood function, equation (3.9), to be re-written as a product of two likelihood terms, L_1 and L_2.
Since $\pi_{000} = 1 - \sum_S \pi_{ijk}$, it follows that (3.9) becomes

$$\underbrace{\begin{pmatrix} N \\ n \end{pmatrix} (\pi_{000})^{N-n} (1-\pi_{000})^{n} \times}_{L_1} \underbrace{\frac{n!}{\prod_S n_{ijk}!} \prod_S \left(\frac{\pi_{ijk}}{1-\pi_{000}} \right)^{n_{ijk}}}_{L_2} \tag{3.10}$$

where S represents the set of all cells apart from the $(0,0,0)$ cell, and remembering that $(1-\pi_{000})^{\sum_S n_{ijk}} = (1-\pi_{000})^n$.
It is clear that L_1 is a binomial function involving the unknown population size, N, and the probability of being missed, p_{000}, and L_2 is the multinomial likelihood giving the conditional distribution for the observed cells.

Evidently, (3.10), is easier to maximize than (3.9) since (3.9) requires simultaneously maximizing with respect to both N and $\{\pi_{ijk}\}$. The conditional maximization carried out in (3.10) finds the $\left(\hat{N}, \hat{\pi}_{000}\right)$ pair that maximizes L_1 and L_2, allowing for the fact that L_1 is conditional on L_2. Given that L_2 effectively precedes L_1, it follows that L_2 needs to be maximized first.

Let the observed probabilities be given by

$$\pi'_{ijk} = \frac{\pi_{ijk}}{1-\pi_{000}}, \tag{3.11}$$

such that $\sum_S \pi'_{ijk} = 1$ where S represents all cells apart from the $(0,0,0)$ cell.
Also let $\mathbf{n}$ represent the observed vector of cell counts $\{n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}\}$. Then the likelihood can be written

$$L\left(N, \pi_{ijk}|\mathbf{n}\right) = L_1\left(N, \pi_{000}|\mathbf{n}\right) \times L_2\left(n, \pi'_{ijk}|\mathbf{n}\right). \tag{3.12}$$

L_2 is multinomial, and can be maximized with respect to the seven observed cell probabilities as

$$\hat{\pi}'_{ijk} = \frac{n_{ijk}}{n} \qquad \text{for all } (i,j,k) \neq (0,0,0). \tag{3.13}$$

In addition, since L_1 is a binomial, it can be maximized with respect to the missing cell probability as

$$\hat{\pi}_{000} = \frac{n_{000}}{N}. \tag{3.14}$$

It becomes clear that (3.14) has three unknowns, but without loss of generality,

$$\hat{\pi}_{000} = \frac{n_{000}}{N} = \frac{\hat{N} - n}{\hat{N}},$$

and after re-arranging this becomes

$$\hat{N} = \frac{n}{(1 - \hat{\pi}_{000})} = \frac{n}{\sum_S \hat{\pi}_{ijk}}. \tag{3.15}$$

The maximum likelihood estimates are computed by first finding the seven $\hat{\pi}'_{ijk}$-terms that maximize L_2. After this has been accomplished, $\hat{N}$ is estimated by maximizing L_1 and making use of (3.11).

Hence, the interpretation of (3.15) is that if an estimate of the missing cell probability can be found (through a model of some sort, or otherwise) then the population size can be subsequently found. This estimator is the same as equation (6.3-6) in Bishop, Fienberg and Holland (1975), with some notational changes. Further, Böhning and Schön (2005) showed that (3.15) is a Horvitz-Thompson type estimator, because of weighting by the inverse of the probability of inclusion.

A remark on the Horvitz-Thompson Estimator

Suppose a sample S of size n is selected from a population, N, such that the probability of the i^{th} being included in the sample is π_i. If the objective is to find the population total, Y then Horvitz and Thompson (1952) found the estimator for the population total to be given by

$$\hat{Y}_{HT} = \sum_S \frac{y_i}{\pi_i} \qquad \text{where } y_i \text{ is a measurement from the } i^{th} \text{ unit.} \tag{3.16}$$

The estimator (3.16) is known as the Horvitz-Thompson estimator, and it has the property of being an unbiased estimator of the population total (Cochran, 1977, page 259). This estimator is important because it represents a milestone in survey methodology, since implicit in their estimator was the idea of sampling weights. In other words, the Horvitz-Thompson estimator can be written as

$$\hat{Y}_{HT} = \sum_S w_i y_i \qquad \text{where } w_i \text{ is the weight associated with unit } i. \tag{3.17}$$

Before the Horvitz-Thompson paper, data collection in sample surveys was done by simple random selection, meaning that every unit in the population had the *same* non-zero chance of inclusion in the sample. The Horvitz-Thompson estimator only requires the inclusion probability of each unit to be non-zero, and thus it allowed for (unequal) probability sampling. With this relaxation of the homogeneous inclusion probability assumption a wide number of sampling schemes (e.g. probability proportional to size, systematic sampling, inverse sampling etc.) were able to be developed. Within a census application the Horvitz-Thompson estimator is important because an adaptation of it is used to produce population estimates based on the post-enumeration survey inclusion probabilities (see Alho (1994) and Brown (2000)).

3.4 The Log-linear Model

The log-linear model is a tool used to make deductions about multi-dimensional contingency tables, and since the seminal paper of Fienberg (1972) has become one of the most widely used techniques in analysing capture-recapture data from multiple lists. The basic premise is that in order to estimate the unknown population size, the general pattern of captures will be represented in the form of a contingency table, and it is therefore possible to model this pattern using the observed individuals. These observed patterns are taken to be a result of some underlying sampling distribution (usually either a Poisson or multinomial), and the log-linear modelling framework allows the exploratory examination and testing of different hypotheses.

The previous Lincoln-Petersen estimator needed to make the explicit assumption of independence, so each individual had the same probability of being observed. However, in reality although this probability is allowed to vary between different samples, the probability of an individual being observed at a certain occasion can be dependent on their past capture history. The log-linear model therefore allows for the relaxation of this independence assumption, especially in the case of multiple lists.

Keeping the same notational format as before, let μ_{ijk} be the expected number of individuals in the $(i, j, k)^{th}$ cell of the 2x2x2 contingency table, with μ_{000} representing the unobserved cell. Also suppose the observed counts n_{ijk} are assumed to have a multinomial distribution with the probabilities associated with each cell given by p_{ijk}. Fienberg (1972) proposed that, for a capture-recapture experiment with r samples, a log-linear model can be selected using the standard hierarchical modelling techniques. Thus in the case of a three-sample experiment, assuming that all cells of the table are fully observed, then the saturated model is

$$\log \mu_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)} + \lambda_{ijk}^{(123)} \tag{3.18}$$

where $\lambda_i^{(1)}$, $\lambda_j^{(2)}$, $\lambda_k^{(3)}$ are the main effect terms,
$\lambda_{ij}^{(12)}$, $\lambda_{ik}^{(13)}$, $\lambda_{jk}^{(23)}$ are the two-way interaction terms,
and $\lambda_{ijk}^{(123)}$ is the three-way interaction term.

The λ term is the normalizing term chosen to make the cell probabilities sum to one. As such,

$$\lambda = -\log \left\{ \sum_i \sum_j \sum_k \exp\left(\lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)} + \lambda_{ijk}^{(123)} \right) \right\}. \tag{3.19}$$

Any solution of (3.18) does not distinguish between the two categories of the cross-classified variables that represent an individual's presence or absence on a given list. Thus in order to be able interpret the model parameters, some constraints are added to (3.19).

These constraints are
$\sum_i \lambda_i^{(1)} = \sum_j \lambda_j^{(2)} = \sum_k \lambda_k^{(3)} = 0$,
$\sum_i \sum_j \lambda_{ij}^{(12)} = \sum_i \sum_k \lambda_{ik}^{(13)} = \sum_j \sum_k \lambda_{jk}^{(23)} = 0$,
and $\sum_i \sum_j \sum_k \lambda_{ijk}^{(123)} = 0$.

The above sum-to-zero constraints effectively treat the parameters symmetrically and allow different combinations of the variable levels to be easily compared. However, an additional constraint is needed for the incomplete 2x2x2 contingency table, with the missing cell. Therefore, the three-way interaction term is set to zero, and the new 'saturated' model in a three-sample capture-recapture experiment becomes

$$\log \mu_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}, \tag{3.20}$$

with constraints
$\sum_i \lambda_i^{(1)} = \sum_j \lambda_j^{(2)} = \sum_k \lambda_k^{(3)} = 0$ and $\sum_i \sum_j \lambda_{ij}^{(12)} = \sum_i \sum_k \lambda_{ik}^{(13)} = \sum_j \sum_k \lambda_{jk}^{(23)} = 0$.

This assumption of the no-three-way interaction is appealing on two fronts. Firstly, it makes intuitive sense under the Bartlett criterion (Bartlett (1935)), and as such the identity

$$\frac{\hat{\mu}_{001}\hat{\mu}_{010}\hat{\mu}_{100}\hat{\mu}_{111}}{\hat{\mu}_{000}\hat{\mu}_{011}\hat{\mu}_{101}\hat{\mu}_{110}} = 1 \quad \text{holds.}$$

Secondly, it becomes possible to define various unsaturated hierarchical models by setting λ-terms in (3.20) to be equal to zero, as shown in Table 3.3. The restriction for all models under consideration to be hierarchical implies that when a particular λ-term is set to zero then all of the higher-order relatives are also zero.

Table 3.3: Log-linear Model Hierarchy

Model	Label	Constraints (terms to be set to zero)
(A)	{123}	None
(B)	{12, 23, 13}	$\{\lambda_{ijk}^{(123)}\}$
(C)	{12, 23}	$\{\lambda_{ijk}^{(123)}, \lambda_{ik}^{(13)}\}$
(D)	{12, 3}	$\{\lambda_{ijk}^{(123)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}\}$
(E)	{12}	$\{\lambda_{ijk}^{(123)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}, \lambda_k^{(3)}\}$
(F)	{1,2,3}	$\{\lambda_{ijk}^{(123)}, \lambda_{ij}^{(12)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}\}$
(G)	{1,2}	$\{\lambda_{ijk}^{(123)}, \lambda_{ij}^{(12)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}, \lambda_k^{(3)}\}$
(H)	{1}	$\{\lambda_{ijk}^{(123)}, \lambda_{ij}^{(12)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}, \lambda_j^{(2)}, \lambda_k^{(3)}\}$
(I)	Constant	$\{\lambda_{ijk}^{(123)}, \lambda_{ij}^{(12)}, \lambda_{ik}^{(13)}, \lambda_{jk}^{(23)}, \lambda_i^{(1)}, \lambda_j^{(2)}, \lambda_k^{(3)}\}$

3.5 The Application of the Log-linear Model to the Census

When dealing with a three-sample census there is a missing cell, n_{000}, which represents the number of individuals absent from the Census, Survey and Third List (see Table 3.4). Therefore, there are only seven observed cells which implies that the 'saturated' model can only have seven parameters.

Table 3.4: Three sample census problem

		Third List			
		Counted		Missed	
		Survey		Survey	
		Counted	Missed	Counted	Missed
	Counted	n_{111}	n_{101}	n_{110}	n_{100}
Census					
	Missed	n_{011}	n_{001}	n_{010}	n_{000}

Chapter 6 Bishop, Fienberg and Holland (1975) provides the derivation of the maximum likelihood estimates under the capture-recapture log-linear framework. The following section summarizes some of the key results, applied in a triple-system setting.

Now for any unsaturated log-linear model, the maximum likelihood estimates for the expected values are given by setting the expected values of the marginal totals corresponding to the highest order λ-terms in the model to be equal to their observed values. It is usual when modelling contingency tables to assume that all the main effects are present, and so the simplest model is the one with mutual independence. Darroch (1958) considered the case where there is sample independence between the three lists, with the corresponding log-linear model becoming

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)}. \tag{3.21}$$

For this model the maximum likelihood estimates for the expected values μ_{ijk} are given by equating the marginal totals corresponding to the highest order terms to their observed values

$$\hat{\mu}_{i++} = n_{i++} \qquad \hat{\mu}_{+j+} = n_{+j+} \qquad \hat{\mu}_{++k} = n_{++k}. \tag{3.22}$$

There is no closed for solution for this independence model since n_{0++}, n_{+0+} and n_{++0} are not observed, but there are some indirect techniques that can be used to find a solution.

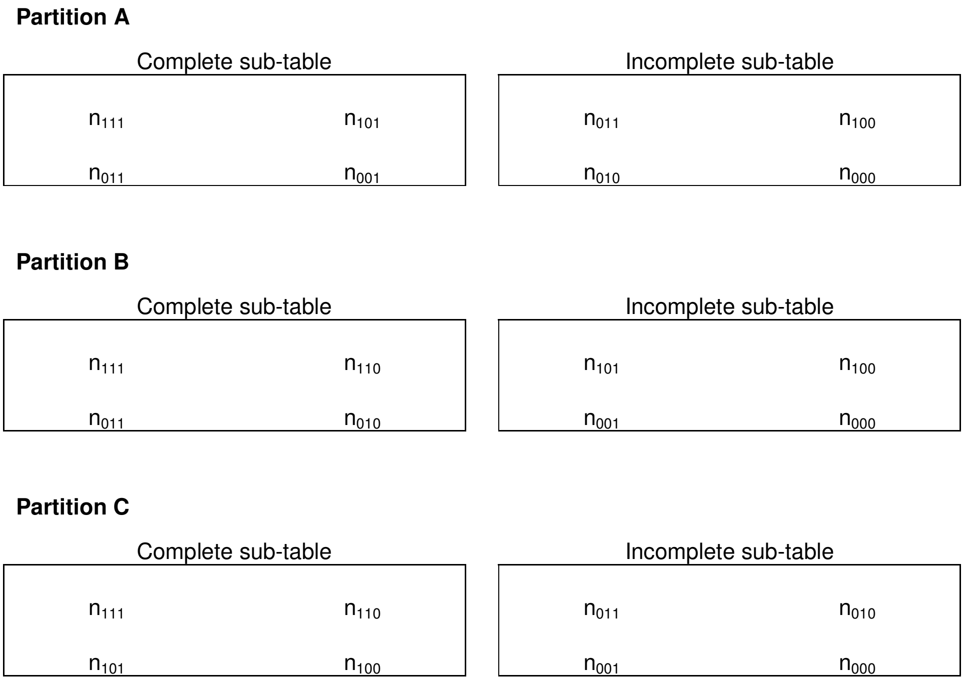
For data presented as an incomplete 2x2x2 contingency table, there are a total of eight possible hierarchical models - the independence model (3.21), three models with a single two-factor interaction term, three with exactly two two-factor interaction terms, and the 'saturated' model with all three two-factor terms. Given that μ_{000} is the expected number of unobserved individuals, then the Bartlett criterion assumption of no three-factor interaction implies that

$$\frac{\mu_{111}\mu_{001}}{\mu_{101}\mu_{011}} = \frac{\mu_{110}\mu_{000}}{\mu_{100}\mu_{010}}. \tag{3.23}$$

It follows that the maximum likelihood estimate for the missing cell count μ_{000} is given by

$$\hat{n}_{000} = \hat{\mu}_{000} = \frac{\hat{\mu}_{111}\hat{\mu}_{001}\hat{\mu}_{100}\hat{\mu}_{010}}{\hat{\mu}_{101}\hat{\mu}_{101}\hat{\mu}_{011}}. \tag{3.24}$$

Figure 3.1: Partitioning of the 2x2x2 contingency table.



Equations (3.23) and (3.24) depend on how the 2x2x2 table is partitioned into two 2x2 sub-tables so that one table is complete, and the other is incomplete as it contains the unobserved cell. Figure 3.1 shows the three ways in which the partitioning could take place. Whichever way the partitioning takes place, it can be seen that the equations used to estimate the missing cell in the form of (3.24) make use of all the available information, and is therefore saturated. The objective is to find the (unsaturated) model with the fewest possible parameters that efficiently accounts for any dependencies that may exist between the different samples. In fact the eight hierarchical models can be summarized into four different models:

1. If the Census, Survey and Third List are mutually independent then

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} \qquad \qquad \textit{Model } \mathbf{I}$$

2. If the Census and Survey are partially independent of the Third List then

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_{ij}^{(CS)} \qquad \qquad \textit{Model } \mathbf{II}$$

3. If the Census and Third List are conditionally independent of each other given the Survey then

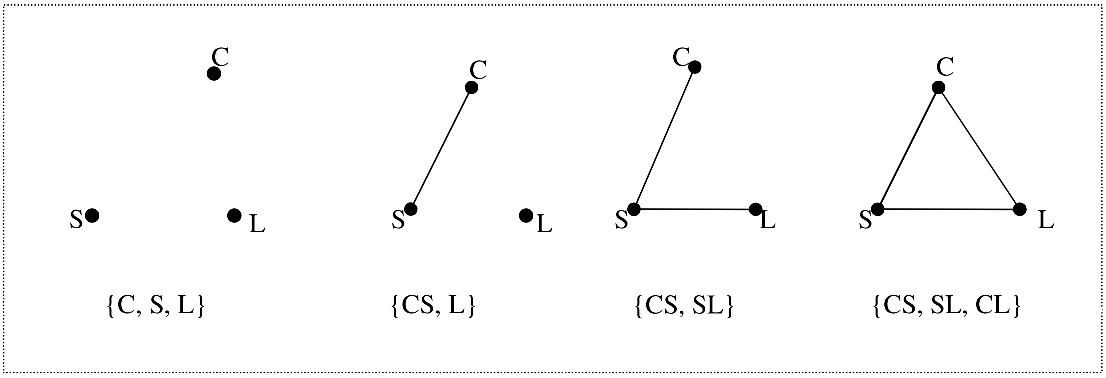
$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_{ik}^{(CS)} + \lambda_{jk}^{(SL)} \qquad \textit{Model } \mathbf{III}$$

4. If the Census, Survey and Third List show pair-wise dependence, in other words there is homogeneous association then

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_{ij}^{(CS)} + \lambda_{ik}^{(CL)} + \lambda_{jk}^{(SL)} \quad \textit{Model } \mathbf{IV}$$

The different log-linear models can be pictorially represented in terms of conditional independence graphs shown in Figure 3.2.

Figure 3.2: Independence Graphs of the Different Log-linear models



When the interaction between the Census and Survey is included then equation (3.21), i.e. Model **II** results, the maximum likelihood equations become

$$\hat{\mu}_{ij+} = n_{ij+} \qquad \text{and} \qquad \hat{\mu}_{++k} = n_{++k}. \tag{3.25}$$

This implies that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}} \qquad \text{where} \quad N = \mu_{+++}. \tag{3.26}$$

In the present format it looks as if (3.25) does not have a direct solution since n_{++0} is not fully known. However, since the Census and Survey are both independent of the Third List it follows that individuals who only appear on the Third List do not provide any information in the estimation of the other observed cells. Therefore, set $\hat{\mu}_{001} = n_{001}$ and if $n' = n - n_{001}$, $n'_{++1} = n_{++1} - n_{001}$ and $n'_{++0} = n_{++0} - n_{000}$ then $\{n_{11+}, n_{10+}, n_{01+}\}$ and $\{n'_{++1}, n'_{++0}\}$ are sufficient statistics for μ_{ij+} and μ_{++k} respectively. As a consequence, the other maximum likelihood estimates can be found by solving the equations

$$\begin{aligned} \hat{\mu}_{ijk} &= \frac{n_{ij+}n_{++1}}{n'} \qquad \text{for } (ijk) \in \{111, 101, 011\} \\ \hat{\mu}_{ijk} &= \frac{n_{ij+}n_{++0}}{n'} \qquad \text{for } (ijk) \in \{110, 100, 010\}. \end{aligned} \tag{3.27}$$

The estimates from (3.27) can be substituted into (3.24) to subsequently get an estimate of the missing cell count n_{000}. It is important to realise that Model **II** is different from the dual system type estimator that is found after summing over the Third List, resulting in cell count n_{11+}, n_{10+}, n_{01+} and n_{00+} (see Table 3.5).

Table 3.5: Contingency table for the DSE, ignoring Third List information

		Second Sample	
		Counted	Missed
	Counted	n_{111}+n_{110}	n_{101}+n_{100}
First Sample			
	Missed	n_{011}+n_{010}	n_{001}+n_{000}

The ensuing DSE under Table 3.5 is given by

$$\frac{n_{00+}n_{11+}}{n_{01+}n_{10+}} = 1 \qquad \Rightarrow \quad \hat{n}_{00+} = \frac{n_{01+}n_{10+}}{n_{11+}}. \tag{3.28}$$

This DSE-type estimator accordingly predicts the margin $\hat{n}_{00+}$, and the missing cell n_{000} is obtained as a consequence by subtraction, since n_{001} is known. This is an important estimator, as it will show how valid the independence assumption is. If the DSE underestimates the population size then the estimate of n_{000} will be negative, due to the fact that the $\hat{n}_{00+}$ is actually less than the observed count added by the Third List, i.e. the n_{001} cell. In Chapter 5, in an application to US Census data, it will be shown how a preliminary analysis of the estimates, $\hat{n}_{00+}$, $\hat{n}_{0+0}$ and $\hat{n}_{+00}$ provides some helpful indications about where there is possible failure in the list independence assumptions.

Under Model **IV**, there is homogeneous association between the Census, Survey and the Third List. Now since the 2x2x2 contingency table can be divided into one complete 2x2 subtable and an incomplete 2x2 subtable, another way of expressing (3.23) is to think of the odds ratio under the complete and incomplete subtables to be equal. Therefore, the odds ratio for the incomplete subtable can be estimated from the complete subtable and the missing cell estimate becomes

$$\hat{n}_{000} = \theta \times \frac{n_{100}n_{010}}{n_{110}} \qquad \text{where} \;\; \theta = \frac{n_{111}n_{001}}{n_{101}n_{011}}. \tag{3.29}$$

It can be seen that when the Census and Survey are conditionally independent of the Third List then $\theta = 1$ (which is Model **III**). Otherwise stated, the maximum likelihood equations have a closed form, and treat $\hat{\mu}_{110} = n_{110}$, $\hat{\mu}_{100} = n_{100}$ and $\hat{\mu}_{010} = n_{010}$ as fixed but solving the remaining terms using

$$\mu_{ijk} = \frac{\mu_{+jk}\mu_{i+k}}{\mu_{++k}}, \tag{3.30}$$

so that $\hat{\mu}_{000} = \frac{n_{+jk}n_{i+k}}{n_{++k}}$.
The missing cell is thus estimated as

$$\hat{\mu}_{000} = \frac{n_{010}n_{100}}{n_{110}}. \tag{3.31}$$

The implication is that under conditional independence the information contained in the complete 2x2 subtable does not play a role in the estimation of the missing cell.

Models **II** and **III** are nested and can therefore be directly compared using the standard chi-squared goodness of fit statistics. In addition, there are other nested versions of the conditional independence and partial independence models (see Figure 3.3). Nonetheless, firstly if the Census and Administrative List could be conditionally independent of the Survey, then the maximum likelihood estimates are derived by fixing
$\hat{\mu}_{001} = n_{001}$, $\hat{\mu}_{100} = n_{100}$ and $\hat{\mu}_{101} = n_{101}$ and the estimates are given by

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{+jk}}{\mu_{+j+}}, \tag{3.32}$$

so that $\hat{\mu}_{ijk} = \frac{n_{ij+}n_{+jk}}{n_{+j+}}$ and the missing cell can be estimated by

$$\hat{\mu}_{000} = \frac{n_{001}n_{100}}{n_{101}}. \tag{3.33}$$

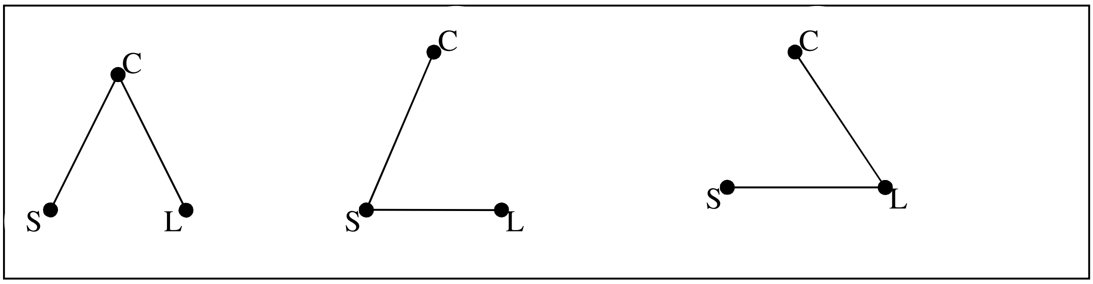
Secondly, the Survey and Administrative List could be conditionally independent of the Census, and the estimated expected values reduce to
$\hat{\mu}_{001} = n_{001}$, $\hat{\mu}_{100} = n_{100}$ and $\hat{\mu}_{011} = n_{011}$,
and

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{i+k}}{\mu_{i++}}, \tag{3.34}$$

with $\hat{\mu}_{ijk} = \frac{n_{ij+}n_{i+k}}{n_{i++}}$ and the missing cell estimate is found by

$$\hat{\mu}_{000} = \frac{n_{001}n_{010}}{n_{011}}. \tag{3.35}$$

Figure 3.3: Independence graphs of the three variants of the conditional independence model.



where C, S and L denote the Census, Survey and Third List.

According to Chapter 6 of Bishop, Fienberg and Holland (1975) the maximum likelihood equation for the case where all the three samples are independent cannot be solved directly. Indirectly, however, the iterative proportional fitting (IPF) algorithm (Deming and Stephan (1940)) can be used to find maximum likelihood estimates of the population total and cell counts, $\hat{N}$ and $\hat{\mu}_{ijk}$. The IPF algorithm works by constraining the marginal totals to the set of minimal sufficient statistics and under the independence model, Model **I**, the cell probabilities satisfy

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}.$$

So the minimum sufficient statistics are $\{\pi_{i++}, \pi_{+j+}, \pi_{++k}\}$ and therefore the iterative algorithm fitted using the observed marginal counts cycles between these three steps

$$\begin{aligned}
\hat{\pi}_{1++} &= \frac{n_{1++}}{\hat{N}}, && (1)\\
\hat{\pi}_{+1+} &= \frac{n_{+1+}}{\hat{N}}, && (2)\\
\hat{\pi}_{++1} &= \frac{n_{++1}}{\hat{N}}. && (3)
\end{aligned}$$

until convergence is reached. Thus, the missing cell when the three lists are independent (i.e. Model **I**) can be estimated in terms of $\hat{N}$ and the marginal probabilities as

$$\hat{n}_{000} = \hat{N}\left(1 - \hat{\pi}_{1++}\right)\left(1 - \hat{\pi}_{+1+}\right)\left(1 - \hat{\pi}_{++1}\right). \tag{3.36}$$

Alternatively, the preferred approach is given by Darroch (1958) who suggested re-expressing the independence between the samples as a quadratic in terms of the sufficient statistics:

$$\left(\hat{N} - n_{1++}\right)\left(\hat{N} - n_{+1+}\right)\left(\hat{N} - n_{++1}\right) = \hat{N}^2\left(\hat{N} - n\right), \tag{3.37}$$

which simplifies to the expression

$$\hat{N}^2\left(n_{1++} + n_{+1+} + n_{++1} - n\right) - \hat{N}\left(n_{1++}n_{++1} + n_{1++}n_{+1+} + n_{+1+}n_{++1}\right) + n_{1++}n_{+1+}n_{++1} = 0.$$

In order to find the best model that fits the data, after the estimated maximum likelihood values $\hat{\mu}_{ijk}$ are known, then the goodness of fit of the models to the observed data can be assessed via the Pearson statistic, X^2, or the log-likelihood ratio statistic G^2, where

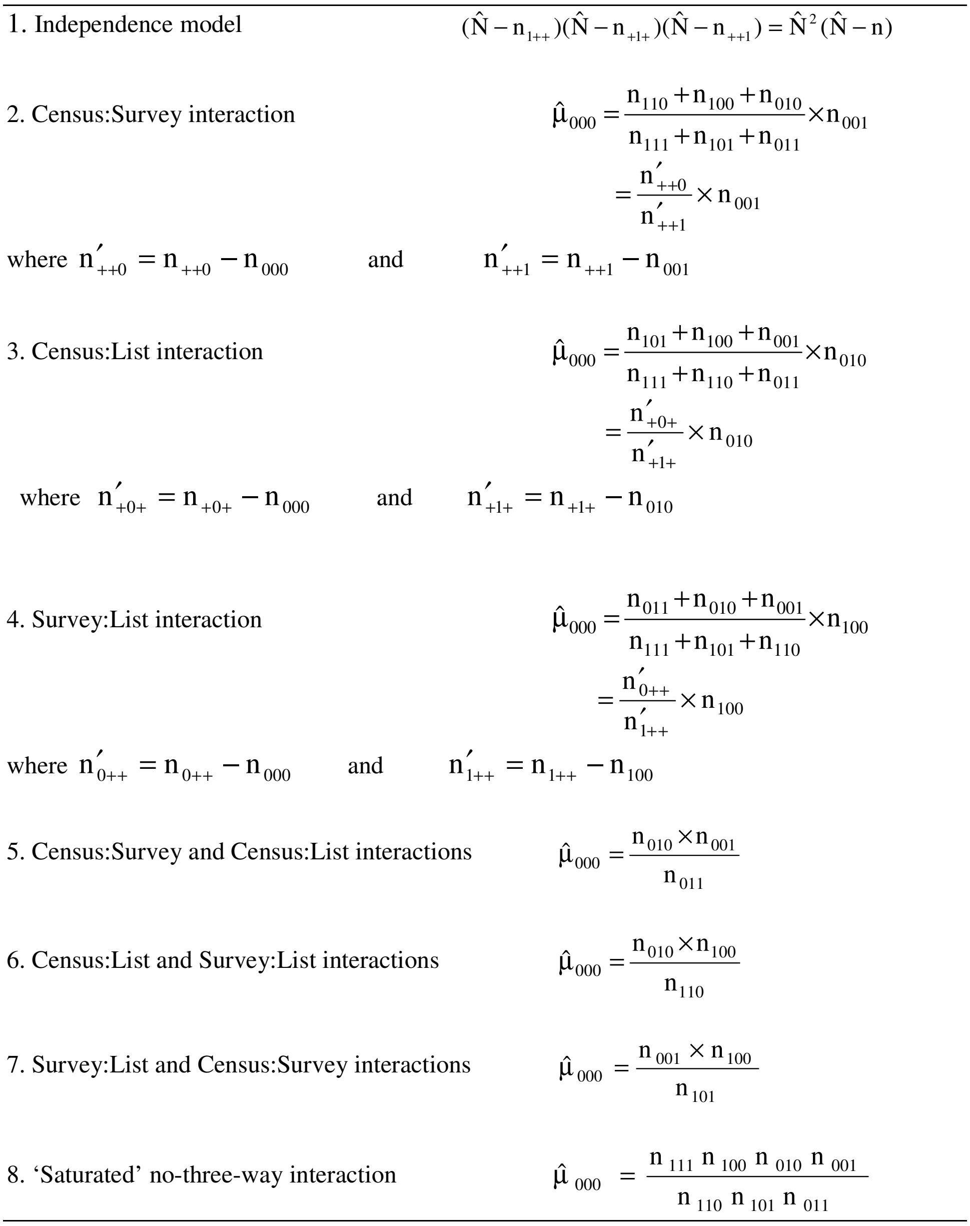
$$\begin{aligned}
X^2 &= \sum_i \sum_j \sum_k \frac{\left(n_{ijk} - \hat{\mu}_{ijk}\right)^2}{\hat{\mu}_{ijk}}\\
G^2 &= 2\sum_i \sum_j \sum_k n_{ijk} \log\left(\frac{n_{ijk}}{\hat{\mu}_{ijk}}\right) \qquad (3.38)\\
&\quad \text{for } (ijk) \neq (000).
\end{aligned}$$

Note that under the chosen model $n_{000} = \hat{\mu}_{000}$.

Since the complete model, with all parameters separately present (i.e. the 'saturated' no-three-way interaction model) is merely a way of re-expressing the data, any of the models proposed is a special case of this 'saturated' model. So if a simpler model truly represents the data, then the difference in Pearson X^2 (or the difference in deviance, calculated using the likelihood ratio G^2) between this model and the more complex model of which it is a special case has approximately χ^2 distribution with degrees of freedom given by the difference in the number of parameters between the models. For the difference between models, the deviance is preferred because it provides a better approximation to the χ^2 and also when a succession of models is being considered, the deviance strictly adds up, while X^2 does not.

The eight log-linear models with the maximum likelihood estimates of the missing cell are given in Table 3.6. In order to make it easier to refer to the models, the table describes each model by their highest order term(s). Incidentally Table 3.6 shows that closed form solutions of the maximum likelihood estimates of the missing cell exist for all the models[1]. Nonetheless, the Expectation Maximization (EM) algorithm will be used to find the missing cell estimates under different models.

Table 3.6: Summary of the missing cell estimates under different log-linear models



1. Independence model	$(\hat{N}-n_{1++})(\hat{N}-n_{+1+})(\hat{N}-n_{++1})=\hat{N}^2(\hat{N}-n)$
2. Census:Survey interaction	$\hat{\mu}_{000}=\dfrac{n_{110}+n_{100}+n_{010}}{n_{111}+n_{101}+n_{011}}\times n_{001}$ $=\dfrac{n'_{++0}}{n'_{++1}}\times n_{001}$
where $n'_{++0}=n_{++0}-n_{000}$ and	$n'_{++1}=n_{++1}-n_{001}$
3. Census:List interaction	$\hat{\mu}_{000}=\dfrac{n_{101}+n_{100}+n_{001}}{n_{111}+n_{110}+n_{011}}\times n_{010}$ $=\dfrac{n'_{+0+}}{n'_{+1+}}\times n_{010}$
where $n'_{+0+}=n_{+0+}-n_{000}$ and	$n'_{+1+}=n_{+1+}-n_{010}$
4. Survey:List interaction	$\hat{\mu}_{000}=\dfrac{n_{011}+n_{010}+n_{001}}{n_{111}+n_{101}+n_{110}}\times n_{100}$ $=\dfrac{n'_{0++}}{n'_{1++}}\times n_{100}$
where $n'_{0++}=n_{0++}-n_{000}$ and	$n'_{1++}=n_{1++}-n_{100}$
5. Census:Survey and Census:List interactions	$\hat{\mu}_{000}=\dfrac{n_{010}\times n_{001}}{n_{011}}$
6. Census:List and Survey:List interactions	$\hat{\mu}_{000}=\dfrac{n_{010}\times n_{100}}{n_{110}}$
7. Survey:List and Census:Survey interactions	$\hat{\mu}_{000}=\dfrac{n_{001}\times n_{100}}{n_{101}}$
8. ‘Saturated’ no-three-way interaction	$\hat{\mu}_{000}=\dfrac{n_{111}\,n_{100}\,n_{010}\,n_{001}}{n_{110}\,n_{101}\,n_{011}}$

[1] Even for the independence model which admittedly does not have a direct estimator for n_{000}, the quadratic estimator is a direct, closed form estimator for N.

3.6 The Expectation Maximization (EM) Algorithm

The EM algorithm is an iterative tool for maximum likelihood estimation in models where there is missing information. When Dempster, Laird and Rubin (1977) introduced the EM algorithm it was designed to be a general iterative procedure for maximum likelihood estimation for incomplete data problems, but the form of this missingness can be in different guises, and that is the appeal of the algorithm. It will be shown during the course of the thesis that the EM algorithm can be adapted to cope with both missingness due to lack of data and the presence of an unobserved variable.

It is composed of two steps, the Expectation (E) step and the Maximization (M) step. The basic idea of the algorithm starts by finding some initial values for the missing parameters that make the log-likelihood less complicated, and effectually easier to maximize. Then this, less complex, log-likelihood is replaced with its expected value calculated at these initial values of the parameters. This is what is known as the **E-step**. Next the modified log-likelihood is maximized, which produces new parameter values. This is known as the **M-step**. The process of alternating between **E** and **M** steps is continued until convergence is reached. Dempster, Laird and Rubin (1977) showed that through the maximizing property of the **M-step** although convergence to the global maximum is not always guaranteed, the log-likelihood will not decrease with each change in the parameters, thereby converging to a (local) maximum. So in order to find the global maximum the algorithm is run from different initial parameter values.

The theory behind the EM algorithm is intertwined with the idea of filling in missing values after making an educated (or otherwise) guess and iterating to find better estimates of the missing values, and Hartley (1958) is one such earlier example which is believed to have laid the ground work for the EM algorithm (see Chapter 8 of Little and Rubin (2002)). In general the framework defines the complete data likelihood to be $\ell\left(\theta|Y_{obs},Y_{mis}\right)$, where the complete data Y is comprised of Y_{obs} and Y_{mis} respectively represents the observed and missing data.

The **E-step**: given an initial estimate $\theta^{(i)}$ of the parameter θ, then

$$\begin{aligned} Q\left(\theta|\theta^{(i)}\right) &= \int_{Y_{mis}} \ell\left(\theta|Y_{obs},Y_{mis}\right) f\left(Y_{mis}|Y_{obs},\theta^{(i)}\right) dY_{mis} \\ &= E\left[\ell\left(\theta|Y_{obs},Y_{mis}\right)|Y_{obs},\theta^{(i)}\right], \end{aligned}$$

i.e. the expected value of the log-likelihood given the observed data, Y_{obs} and the parameter estimate, $\theta^{(i)}$.

The **M-step**: maximize $Q\left(\theta|\theta^{(i)}\right)$ to obtain the next iterate, $\theta^{(i+1)}$ such that

$$Q\left(\theta^{(i+1)}|\theta^{(i)}\right) \geq Q\left(\theta|\theta^{(i)}\right).$$

Under most cases, the log-likelihood function is linear in Y_{mis}, and so continued updating of the parameter value at the E and M steps will lead to the maximum likelihood estimate, $\hat{\theta}$. The convergence is reached when $|Q\left(\theta^{(i+1)}|\theta^{(i)}\right) - Q\left(\theta|\theta^{(i)}\right)|$ is small.

In order to heuristically prove that the E and M steps do iteratively converge to a solution, it must be noted that the complete data can be factorized as

$$f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta) = f(Y_{obs}|\theta)\, f(Y_{mis}|Y_{obs}, \theta)$$

and the log-likelihood can be decomposed as

$$\ell(\theta|Y) = \ell(\theta|Y_{obs}, Y_{mis}) = \ell(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs}, \theta).$$

It is required to find an estimate of θ that maximizes the incomplete data log-likelihood, $\ell(\theta|Y_{obs})$, which is not easy to directly do. However, it may be noticed that incomplete data log-likelihood can be re-written as

$$\ell(\theta|Y_{obs}) = \ell(\theta|Y) - \log f(Y_{mis}|Y_{obs}, \theta). \tag{3.39}$$

The complete data log-likelihood, $\ell(\theta|Y)$, and the missing part of the complete data log-likelihood, $\log f(Y_{mis}|Y_{obs}, \theta)$, in (3.39) are relatively easy to maximize. Therefore, taking the expectation of both sides of (3.39) over Y_{mis} given the Y_{obs} and a current estimate of θ, $\theta^{(i)}$ yields

$$\ell(\theta|Y_{obs}) = Q\left(\theta|\theta^{(i)}\right) - H\left(\theta|\theta^{(i)}\right),$$

where

$$Q\left(\theta|\theta^{(i)}\right) = \int_{Y_{mis}} \ell(\theta|Y_{obs}, Y_{mis})\, f\left(Y_{mis}|Y_{obs}, \theta^{(i)}\right) dY_{mis}$$

and

$$\begin{aligned} H\left(\theta|\theta^{(i)}\right) &= \int_{Y_{mis}} \left[\log f(Y_{mis}|Y_{obs}, \theta)\right] f\left(Y_{mis}|Y_{obs}, \theta^{(i)}\right) dY_{mis} \\ &= E\left[\log f(Y_{mis}|Y_{obs}, \theta)\,|Y_{obs}, \theta^{(i)}\right]. \end{aligned}$$

At successive iterates,

$$\ell\left(\theta^{(i+1)}|Y_{obs}\right) - \ell\left(\theta^{(i)}|Y_{obs}\right) = \left[Q\left(\theta^{(i+1)}|\theta^{(i)}\right) - Q\left(\theta^{(i)}|\theta^{(i)}\right)\right] - \left[H\left(\theta^{(i+1)}|\theta^{(i)}\right) - H\left(\theta^{(i)}|\theta^{(i)}\right)\right].$$

The maximizing property of the EM algorithm is that if $\theta^{(i+1)}$ is chosen such that $Q\left(\theta^{(i+1)}|\theta^{(i)}\right)$ is greater than $Q\left(\theta^{(i)}|\theta^{(i)}\right)$, then firstly the differences of the Q functions are always positive, and secondly the differences in H functions are negative[2]. Hence, for any EM algorithm successive iterative changes from $\theta^{(i)}$ to $\theta^{(i+1)}$ leads to an increase in the log-likelihood; i.e. $\ell\left(\theta^{(i+1)}|Y_{obs}\right) - \ell\left(\theta^{(i)}|Y_{obs}\right) \geq 0$. A more rigorous proof is presented in Dempster, Laird and Rubin (1977).

[2] Using Jensen's inequality $H\left(\theta|\theta^{(i)}\right) \leq H\left(\theta^{(i)}|\theta^{(i)}\right)$.

3.7 Variance Estimation

Since the population size, N, is an estimate there is anticipated to be some variability involved. Although there may be some statistic which asserts that one of the models is marginally better than the others, it is worthwhile to quote not just the best-fitting model and its estimate of the population size but an estimate of its precision, as well. There are a number of approaches for computing confidence intervals for functions of maximum likelihood estimates, and this section will be considering four of them that were used in the thesis - the profile likelihood-based, the Delta Method-based, Bootstrap resampling and the Supplemented EM algorithm-based confidence intervals.

The profile likelihood confidence intervals are based on the asymptotic χ^2 distribution of the generalized likelihood ratio test and are better behaved that the Wald confidence intervals particularly when there are small sizes (Evans et al. (1996)). On the other hand, the Delta Method and Supplemented EM (SEM) algorithm both produce asymptotic variances and are therefore only guaranteed to be inferentially valid under asymptotic conditions. However, while the covariance-variance matrix obtained using the SEM is based on the observed second derivatives of the observed data likelihood, the Delta Method uses a Taylor Series approximation to first expand a function of a random variable about its mean, and then take the variance of this expanded function. The bootstrap does not make any distributional assumptions about the sampling distribution, and can therefore produce better confidence intervals, particularly when there is evidence of some skew to the data.

3.7.1 Profile Likelihood

In a capture-recapture context Agresti (1994) and Cormack (1992) use the profile likelihood function, where the profile likelihood writes the likelihood as a function of the unobserved cell count, n_{000}. Their approach is to construct confidence intervals for N on the basis of its profile likelihood function found by substituting different values of n_{000} then estimating the model parameters by maximizing the likelihood function - in effect the likelihood is 'profiled' over n_{000}-values. The deviance can then be evaluated, and the confidence limits are the values of $N = n + n_{000}$ that yield a deviance differing from the prescribed $\chi^2_1(\alpha)$ value, here at the 95% confidence interval this value is 3.84. The motivation behind the use of the profile likelihood is that profile likelihood confidence intervals often have better small-sample properties than those based on asymptotic standard errors calculated from the full likelihood (Cormack (1992)). Also, unlike conditional and marginal likelihoods, profile likelihood methods can always be used (even when the profile likelihood cannot be written down explicitly, there are numerical methods that can be used). However, the computation of profile likelihood confidence intervals is not simple and often requires time-consuming optimization procedures.

The computation of the profile likelihood can be difficult due the need to search over all possible values that would lead to the rejection of the null hypothesis, i.e. yield non-significance. However, most computer packages offer profile likelihood confidence intervals within their modelling procedures. In the **R** environment, there is a special function *plkhci*, based on Venzon and Moolgavkar's algorithm (Venzon and Moolgavkar (1988)) available in the special add-on package Bhat which works by inverting the likelihood ratio test statistics. *plkhci* requires the negative log-likelihood which for the multinomially distributed contingency table is

$$\begin{aligned} -\log L &= -\frac{N!}{\prod_i \prod_j \prod_k n_{ijk}!} \sum_i \sum_j \sum_k n_{ijk} \log \pi_{ijk} \\ &= -\left\{ \log N! - \log(N-n)! - \sum_S n_{ijk}! + n_{000} \log\left(\frac{N-n}{N}\right) + \sum_S \frac{n_{ijk}}{N} \right\} \end{aligned}$$

where, as per usual, S represents the set of all cells apart from the $(0,0,0)$ cell and n is the sum of the observed cells.

This likelihood is still not easy to maximize owing in part to the combinatorial term $\begin{pmatrix} N \\ (n_{ijk}) \end{pmatrix}$ since N is unknown, and as such the conditional likelihood is used which is given by

$$L_c\left(\pi_{ijk}; n_{obs}|n\right) \propto \prod_i \prod_j \prod_k \left(\pi_{ijk}^{'}\right)^{n_{ijk}}.$$

This yields the negative conditional log-likelihood,

$$\log L_c \propto \sum_i \sum_j \sum_k n_{ijk} \log \pi_{ijk}^{'}$$

where

$$\pi_{ijk}^{'} = \frac{\pi_{ijk}}{\sum_S \pi_{ijk}} \Rightarrow \hat{\pi}_{ijk}^{'} = \frac{\frac{n_{ijk}}{N}}{\sum_S \frac{n_{ijk}}{N}}.$$

3.7.2 Delta Method

The Delta method is used to derive an approximate probability distribution for a function of an asymptotically normal statistical estimator based on knowledge of the limiting variance of that estimator. In essence, the Delta method takes a function whose variance cannot be analytically computed as it is deemed to be too complex. Using results from the Central Limit Theorem, the Delta method creates a linear approximation of that complex function and then computes the variance of the simpler linear function which can then be used for large sample inference. The Delta method is therefore a technique for deriving standard errors for large sample inference based on finding asymptotic distribution of the parameter of interest - note that from the Central Limit Theorem the limiting behaviour of parameter is the asymptotic normal distribution. Bishop, Fienberg and Holland (1975) derived the asymptotic variance of the population estimate for the different log-linear models using the Delta method; these are summarized below.

When the three samples are mutually independent (i.e. Model **I**), the variance is estimated as

$$\hat{V}\left(\hat{N}\right) = \frac{\hat{N}\hat{\mu}_{000}}{n_{011} + n_{101} + n_{110} + n_{111}}. \tag{3.40}$$

The asymptotic variance for the model with one pair-wise interaction (i.e. Model **II**) is estimated as

$$\hat{V}\left(\hat{N}\right) = \left(\hat{\mu}_{000}\right)^2 \left(\frac{1}{n_{++1} - n_{001}} + \frac{1}{\hat{n}_{++0} - \hat{n}_{000}} + \frac{1}{n_{001}} + \frac{1}{\hat{\mu}_{000}}\right). \tag{3.41}$$

When there is conditional independence, and assuming Model **III** holds, i.e. the Census and Survey are conditionally independent of the Third List, then the asymptotic variance is estimated as

$$\hat{V}\left(\hat{N}\right) = \left(\hat{\mu}_{000}\right)^2 \left(\frac{1}{n_{001}} + \frac{1}{n_{010}} + \frac{1}{n_{011}} + \frac{n_{011}}{n_{001}n_{010}}\right). \tag{3.42}$$

Finally for the saturated model, the asymptotic variance estimate is estimated as

$$\hat{V}\left(\hat{N}\right) = \left(\hat{\mu}_{000}\right)^2 \left(\frac{1}{n_{111}} + \frac{1}{n_{110}} + \frac{1}{n_{101}} + \frac{1}{n_{100}} + \frac{1}{n_{011}} + \frac{1}{n_{010}} + \frac{1}{n_{001}} + \frac{1}{\hat{\mu}_{000}}\right). \tag{3.43}$$

In addition the dual system estimate asymptotic variance was derived by Chandrasekar and Deming (1949), also using the Delta method, and is given by

$$\hat{V}\left(\hat{N}\right) = \frac{n_{+1}n_{1+}n_{01}n_{10}}{\left(n_{11}\right)^3}. \tag{3.44}$$

In analysing capture-recapture three-sample data the aim is to fit the incomplete 2x2x2 table by a log-linear model with the fewest possible parameters. It becomes readily apparent that the fewer the parameters in the 'most suitable' model for estimating μ_{000} the smaller the variance. Thus it becomes ideal not to just use the 'saturated model' given in (3.20). On the other hand if a model with too few parameters is used, bias may be introduced into the resulting estimate of the population size, and the risk is that the variance formulae (3.40)-(3.43) become meaningless. Essentially, the variance formulae only hold under the assumption that the correct model has been chosen.

3.7.3 Supplemented Expectation Maximization Algorithm (SEM)

The Supplemented EM algorithm is an extension of the EM algorithm that facilitates the calculation of the variance-covariance matrix, under large sample conditions. Now, it is immediate that the information matrices associated with the complete data likelihood from the EM algorithm do not directly yield valid asymptotic covariance estimates for the estimated parameters. This is because there is an additional component that is required to account for the influence of the missingness. Therefore, the SEM algorithm finds the standard errors of the maximum likelihood estimates, but has the advantage over other techniques because it does not require the computing and inverting of the information matrix; the difficulty of likelihood-based methods lies in the computation of the second derivative, and then taking the inverse in order to obtain the information matrix.

The SEM algorithm was proposed by Meng and Rubin (1991) where they calculate the large-sample variance-covariance matrix associated with the parameter estimates (under maximum likelihood estimation) using

i. the E and M steps of the EM algorithm,
ii. the large-sample complete data variance-covariance matrix, and
iii. standard matrix operations.

Thus under the SEM algorithm, the desired observed data variance-covariance matrix, V_{obs}, can be derived from the complete data variance-covariance matrix, V_{com} and a matrix DM, which is determined by the rate of convergence of the EM algorithm. Now define the complete information to be i_{com}, and the observed and missing information to be i_{obs} and i_{mis} respectively, then it follows that

$$i_{com} = i_{obs} + i_{mis}. \tag{3.45}$$

Also define the fraction of the missing information to be DM, in other words,

$$DM = i_{mis} i_{com}^{-1} = I - i_{obs} i_{com}^{-1}. \tag{3.46}$$

Note that equation (3.46) is found by simply rearranging (3.45) and 'dividing' by i_{com}. An associated result (from Dempster, Laird and Rubin (1977)) is that this matrix, DM, is the gradient of the EM mapping and controls the EM algorithm's speed of convergence; inherently, the larger the fraction of the missingness, the slower it will take to reach convergence.

Equation (3.46) implies that

$$\begin{aligned} i_{obs}^{-1} &= i_{com}^{-1}[I - DM]^{-1} \\ \Rightarrow \quad V_{obs} &= V_{com}[I - DM]^{-1}. \end{aligned} \tag{3.47}$$

Consequently, (3.46) can be re-written as

$$V_{obs} = V_{com}\left(I - DM + DM\right)\left(I - DM\right)^{-1} = V_{com} + \Delta V,$$

where $\Delta V = V_{com} DM \left(I - DM\right)^{-1}$ is the increase in variance due to the missing data.

Recall that the EM algorithm works by defining a mapping, M given by $\theta^{(t+1)} \rightarrow M\left(\theta^{(t)}\right)$, which will converge so that $\theta^{*} \rightarrow M\left(\theta^{*}\right)$. Then DM is the Jacobian matrix where

$$DM = \left(\frac{\partial M_j\left(\theta\right)}{\partial \theta}\right)|_{\theta=\theta^{*}},$$

which can be numerically estimated.

The key idea of the SEM algorithm is that even though the mapping, M, does not have an explicit mathematical form, its derivative, DM can be estimated as a by-product of the EM calculations. This is due to a special feature of the EM algorithm that means that the derivative of the mapping, DM, is defined by the EM steps; therefore, the numerical differentiation is carried out automatically between iterations. Furthermore, despite there being other techniques available to obtain the variance-covariance matrix based on re-sampling from the empirical distribution (such as the bootstrap or the jack-knife), the fact that these re-sampling techniques work best with large samples with independent and identically distributed structures can be a limitation, particularly in cases where there are complicated missing data patterns. Since the SEM algorithm obtains the desired variance-covariance matrix by modifying the complete data variance-covariance matrix with an increment due to the missing data, and makes no additional assumptions regarding the data structure, the resulting variance-covariance matrix is a much better asymptotic approximation than that obtained using the jack-knife or bootstrap (Meng and Rubin (1991)). Van Deusen (2002), in an application to capture-recapture where a logistic model was fitted to an open population, compared the estimates of the parameters under the EM algorithm and the Newton-Raphson iterative procedure and applied the SEM algorithm to obtain the asymptotic variance-covariance matrix of the parameters.

3.7.4 Bootstrap

The bootstrap is a computationally intensive procedure that relies on resampling from the observed data and can be used to determine biases, standard errors, confidence intervals, amongst other statistical parameters of interest, in circumstances where it is not easy to theoretically obtain these statistics. It was first introduced by Efron (1979), and the theoretical ideas more clearly formalised in Efron and Tibshirani (1993). The basic idea behind the bootstrap is to estimate the sampling distribution of a parameter by resampling. The concept of the bootstrap is similar to the, more established, jack-knife. However, unlike the jack-knife which is mostly concerned with calculating standard errors of parameters, the bootstrap estimates not only the standard errors but also the distribution of the estimator of interest.

In statistical theory, the sampling distribution is usually derived from random sampling from the population a number of times and through the Central Limit Theorem inferences may be made from the sample about the population. In bootstrapping, instead of taking samples from the population, *resamples* are created by repeated sampling, *with replacement* from the initial observed sample that is the same size as the original observed sample. The bootstrap distribution of the resamples is used to estimate how the initial observed sample varies due to random sampling. As such the bootstrap procedure is used to first, estimate the parameter of interest and second, estimate the variability of the parameter estimate. This is done without recourse to the Central Limit Theorem, and is therefore the main advantage of the bootstrap over other precision estimation techniques.

In Efron's 1979 paper, he distinguishes between the *parametric* and *non-parametric* bootstrap, and the choice of bootstrapping procedure is dependent on whether parametric or non-parametric inference is being made. In bootstrapping the unknown distribution, F, is estimated by the empirical distribution, $\hat{F}$, which is found by resampling from the original sample. Therefore for the non-parametric version, each sample item is assigned equal selection probability. On the other hand for the parametric version, the unknown distribution is considered to be from a prescribed parametric family. For data collected by capture-recapture a parametric bootstrap is preferred to a non-parametric bootstrap (Buckland and Garthwaite (1991)), and in the parametric bootstrap implemented in the course of the thesis the data are assumed to be from the multinomial distribution.

3.8 A Definition of Identifiability

In statistical modelling, the belief is that the underlying statistical distribution is completely known, with the exception of a set of unknown parameters. Hence, the primary objective of modelling is to find (or estimate) these unknown parameters. But it is entirely feasible that several sets of unknown parameters could have generated the underlying statistical distribution. As such, the problem of *identifiability* of the model parameters basically concerns whether or not the values of the parameters are uniquely determined by the observed data.

A more formal definition, based on Chapter 5 of Skrondal and Rabe-Hesketh (2004), states that given a model with likelihood function $\mathbf{L}(\vartheta)$ then the model is not identified if for $\vartheta_1 \neq \vartheta_2$, $\mathbf{L}(\vartheta_1) = \mathbf{L}(\vartheta_2)$. By defining identifiability in terms of the likelihood function and noting that the likelihood function depends on the *information* the observed data carry about the unknown parameters, it becomes clear that the lack of identifiability translates into a lack of sufficient information. In addition, there is a link between the variance $V\left(\hat{\vartheta}\right)$, information $\mathbf{I}\left(\hat{\vartheta}\right)$ and likelihood $\mathbf{L}(\vartheta)$ where

$$\mathbf{I}(\vartheta) = -\mathrm{E}\left[\frac{\partial^{\mathbf{2}}}{\partial\vartheta^{\mathbf{2}}}\log\mathbf{L}(\vartheta)\right]$$

and

$$V\left(\hat{\vartheta}\right) = \frac{1}{\mathbf{I}(\vartheta)}.$$

Therefore, in order to be able to compute the variance of a parameter estimate the inverse of the information needs to be taken, and as a corollary for matrices if the information matrix is non-invertible, the variance cannot be computed. Rothenberg (1971) made the connection between non-invertibility of the information matrix and identifiability.

The notion of weak identifiability is much more difficult to formally define (Gelfand and Sahu (1999)). But simply stated, weak identifiability occurs when the data supplies little information about some of the parameters. Under the Bayesian paradigm, Gelfand

and Sahu (1999) say that a parameter ϑ_2 is weakly identified if there exists ϑ_1 such that the posterior distribution, $f(\vartheta_2|\vartheta_1, y)$ is roughly equivalent to the the prior distribution, $f(\vartheta_2|\vartheta_1)$. So,

$$f(\vartheta_2|\vartheta_1, y) \approx f(\vartheta_2|\vartheta_1) \qquad \text{where } y \text{ is the data.}$$

However they argue, using a result from Dawid (1979), that Bayesian non-identifiability is equivalent to the lack of identifiability in the likelihood. This, therefore, implies that identifiability does not depend upon the specification of the prior distribution. In other words, when the likelihood does not contain information with regards to some parameters, the nature of the prior specification drives the posterior results. A model that does not contain information on some parameters is said to have a flat likelihood (Lindley and Smith (1972)). A flat likelihood function means that all possible values of the parameter are almost equally likely. So in the classical paradigm, maximum likelihood estimation when there is a flat likelihood can be difficult - for instance, the EM algorithm converges extremely slowly. On the other hand, the maximum likelihood estimation for a peaked likelihood function is much simpler and the results are generally more precise (i.e. have smaller standard errors). For complex models such as the latent class models being proposed in this thesis, identifiability - and specifically weak identifiability - will be shown to be an issue.

3.9 Latent Class Analysis

The aim of latent class analysis is to define a latent variable as a set of classes within which the manifest variables are locally independent. Latent class models rely on two central assumptions. The first one is that the population consists of a set of internally homogeneous and mutually exclusive subpopulations, which make up a latent classification that is discrete by nature. The other is local independence, which means that within a given latent subpopulation, all the manifest variables are statistically independent. In other words, the manifest variables are conditionally independent given the categories of the latent variable. The latent class analysis can either be exploratory or confirmatory. In the former case there are no a priori restrictions on the parameters of the model, whereas in the latter case some restrictions can be imposed.

In order to evaluate the coverage error in a population census, many countries conduct a post-enumeration survey which is designed to identify individuals who were missed in the census, as well as individuals who were counted in the census, but should not have been. The quality of this evaluation process relies heavily on the ability of the survey to accurately classify these erroneous enumerations. In previous UK censuses, the number of erroneous enumerations has been assumed to be negligible in comparison with the underenumeration. This is not necessarily true; as the previous sections have highlighted that the dual system estimator of the population can be susceptible to some biases arising from heterogeneous enumeration probabilities and the lack of independence in enumeration error between the Census and the Survey. Therefore, latent class analysis has been suggested (for example by Biemer et al. (2001b)) as a way of identifying the individuals counted in the census process by their true residence status.

In a census enumeration context, let X_p denote a dichotomous variable defined for the p^{th} person in the population, where

$$X_p = \begin{cases} 1 & \text{if individual } p \text{ is an actual enumeration;} \\ 0 & \text{if individual } p \text{ is an erroneous enumeration.} \end{cases}$$

The assumption is that X_p is unknown and unobservable - and is therefore a latent variable. However, there are some observable indicators that can serve as proxies of X_p. Biemer et al. (2001a) use the Census, the Survey and data from administrative records as a Third List. On the other hand, Biemer et al. (2001b) use the reconciled re-interview of survey respondents as the Third List. In both cases there are three indicators of X_p which correspond to enumeration in the census, denoted by C_p, enumeration in the coverage survey, denoted by S_p, and enumeration on the third list, denoted by L_p. Like the latent variable X_p, each of these indicators are dichotomous, taking the value 1 if individual p is counted and 0 if missed. For brevity, the subscripts in (C_p, S_p, X_p, L_p) are dropped and the data can be represented in the form of the contingency table, in Table 3.7, with observed cell counts n_{ijkt}.

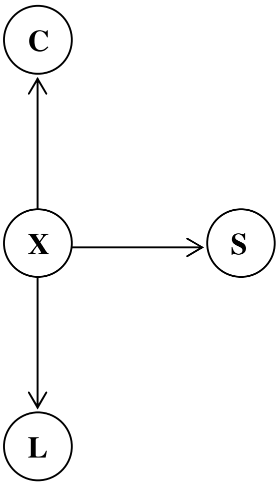
Table 3.7: Contingency table with a latent variable

Latent Class 1 - Real					
		Third List			
		Counted		Missed	
		Survey		Survey	
		Counted	Missed	Counted	Missed
Census	Counted	n_{1111}	n_{1011}	n_{1101}	n_{1001}
	Missed	n_{0111}	n_{0011}	n_{0101}	n_{0001}
Latent Class 2 - Erroneous					
		Third List			
		Counted		Missed	
		Survey		Survey	
		Counted	Missed	Counted	Missed
Census	Counted	n_{1112}	n_{1012}	n_{1102}	n_{1002}
	Missed	n_{0112}	n_{0012}	n_{0102}	n_{0002}

There are two parameterizations of the latent class model. One is the classical parameterization, developed by Lazarsfeld and Henry (1968) and Goodman (1974), in which the model parameters are the latent class prevalences and conditional response probabilities. The other reparameterizes the latent class model as a log-linear model, and was developed by Haberman (1979). As a result of the ease with which the log-linear models, considered earlier in Section 3.5, can be expanded to include latent variables the thesis will focus on the Haberman parameterization. In both parameterizations, Figure 3.4 shows the relationship between the Census (C), Survey (S) and Third List (L) indicators and the latent variable (X).

In this section, for the time being, it is assumed that the 2x2x2 table representing the capture histories in the Census, Survey and Third List is fully observed, i.e. n_{000} is known. The generalisation, with n_{000} unknown, is considered in Sections 3.10 and 3.11.

Figure 3.4: Independence graph showing the relationship between the latent and manifest variables



To specify the latent class model, let π_{ijk} denote the probability that an individual will be at level (i, j, k) with respect to the joint variable (C, S, L) for $i = 0, 1$; $j = 0, 1$; $k = 0, 1$; and let π_{ijkt}^{CSLX} represents the probability that a randomly chosen individual will be in the $(i, j, k, t)^{th}$ cell of the joint variable with (C, S, L, X) for the levels of the latent variable given by $t = 1, 2$. Also let π_t^X denote the latent class probability, that is the probability that an individual will be at level t with respect to the latent variable. Finally, let $\pi_{it}^{C|X}$, $\pi_{jt}^{S|X}$ and $\pi_{kt}^{L|X}$ be the conditional probabilities, where $\pi_{it}^{C|X}$ denotes the conditional probability that an individual will be at level i with respect to the Census variable, given that they are at level t of the latent variable X, $\pi_{jt}^{S|X}$ denotes the conditional probability that an individual will be at level j with respect to the Survey variable, and $\pi_{kt}^{L|X}$ denotes the conditional probability that an individual will be at level k with respect to the Third List. Then under the conditional probabilities parameterization,

$$\pi_{ijkt}^{CSLX} = \pi_t^X \pi_{it}^{C|X} \pi_{jit}^{S|CX} \pi_{kijt}^{L|CSX}$$

which is equivalent to

$$\pi_{ijkt}^{CSLX} = \pi_t^X \pi_{it}^{C|X} \pi_{jt}^{S|X} \pi_{kt}^{L|X}. \tag{3.48}$$

These conditional probabilities given in equation (3.48) are defined in Table 3.8.

The implication is that the probability a randomly selected case will be located in cell (i, j, k, t) can be decomposed into the product of the appropriate marginal and conditional probabilities. When the classifiers, C, S, L are assumed to be mutually independent then

Table 3.8: Conditional Probabilities parameterization

	Census		Survey		Third List	
	Counted	Missed	Counted	Missed	Counted	Missed
Class 1	$\pi_{11}^{C\|X}$	$\pi_{01}^{C\|X}$	$\pi_{11}^{S\|X}$	$\pi_{01}^{S\|X}$	$\pi_{11}^{L\|X}$	$\pi_{01}^{L\|X}$
Class 2	$\pi_{12}^{C\|X}$	$\pi_{02}^{C\|X}$	$\pi_{12}^{S\|X}$	$\pi_{02}^{S\|X}$	$\pi_{12}^{L\|X}$	$\pi_{02}^{L\|X}$

this equals the product of the probability of a randomly selected case being at level t of the latent variable X times the conditional probabilities that an individual in class t of the latent variable will be located in a certain category of each of the manifest variables. So simply put, within a latent class, probabilities multiply, as shown in equation (3.48).

The conditional probabilities represent a measure of the degree of association between each of the manifest variables and each of the latent classes and can be compared to the factor loadings in a factor analysis. Just like in factor analysis, the latent variables are at their most useful in terms of explaining the relationships between the manifest variables if they have some theoretically meaningful interpretation.

The likelihood can be written as

$$\mathbf{L} = \prod_i \prod_j \prod_k \prod_t \left(\pi_{ijkt}^{CSLX}\right)^{n_{ijkt}} \tag{3.49}$$

and the log-likelihood is

$$\ell = \sum_i \sum_j \sum_k \sum_t n_{ijkt} \log \pi_{ijkt}^{CSLX}. \tag{3.50}$$

Haberman (1979) showed that (3.48) is equivalent to the log-linear model

$$\log \mu_{ijkt} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)}, + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} \tag{3.51}$$

where the μ_{ijkt} are the expected counts in the $(i, j, k, t)^{th}$ cell.
The relationship between the two formulations of the latent class models can be illustrated by writing the conditional probabilities in (3.48) as a function of the log-linear parameters appearing in equation (3.51). Thus for a example, when the manifest variables are binary (with categories 0 and 1) and the model posits two latent classes (with classes 1 and 2) then the conditional probability that the individual is missed in the Census given that they are in the first class is

$$\Pr(C = 0|X = 1) = \pi_{01}^{C|X} = \frac{\exp\left(\lambda_0^{(C)} + \lambda_{01}^{(CX)}\right)}{\exp\left(\lambda_0^{(C)} + \lambda_{01}^{(CX)}\right) + \exp\left(\lambda_1^{(C)} + \lambda_{11}^{(CX)}\right)}.$$

It has to be stated here that the latent class framework cannot be applied to the case with two manifest variables (i.e. dual system estimation). The proof is as follows. Denoting the latent variable as L, in the usual notation, then the latent class model with two manifest variables is given by

$\log \mu_{ijt}^{(CSX)} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)}$ under the Haberman parameterization, or

$$\pi^{CSX}_{ijt} = \pi^X_t \pi^{C|X}_{it} \pi^{S|X}_{jt} \qquad \text{under the Goodman parameterization.}$$

Since there are five free parameters[3] to be estimated, namely $\{\lambda^{(C)}_1, \lambda^{(S)}_1, \lambda^{(X)}_1, \lambda^{(CX)}_{11}, \lambda^{(SX)}_{11}\}$ using the Haberman parameterization or $\{\pi^X_t, \pi^{C|X}_{0t}, \pi^{C|X}_{1t}, \pi^{S|X}_{0t}, \pi^{S|X}_{1t}\}$ using the Goodman parameterization, but four cell probabilities (three knowns because the probabilities should sum to one), the number of degrees of freedom is not non-negative, and thus the latent class model is not identified. Some constraints may be imposed on the model, such as $\lambda^{(C)}_i = \lambda^{(S)}_j$ and $\lambda^{(CX)}_{it} = \lambda^{(SX)}_{jt}$ (or equivalently $\pi^{C|X}_{it} = \pi^{S|X}_{jt}$).

The interpretability of the ensuing model may be difficult, as well as being unrealistic. However, when there are three manifest variables then the model is just identified with zero degrees of freedom, since there are now seven parameters and eight cell probabilities (but seven knowns since the probabilities must sum to one). So it follows that the lowest number of manifest variables possible for the identification of latent class models without resorting to imposing restrictions is three.

3.9.1 Maximum Likelihood Estimation of Parameters under the Goodman Parameterization for the Local Independence Latent Model

Maximum likelihood estimation under the two latent class parameterizations (i.e. the Goodman conditional probabilities and Haberman log-linear models) will now be presented in greater detail. It must be noted that the contingency table representing the capture histories is assumed to have no missingness, in other words n_{000} is known. However, the extension of the estimation process when there is both latentness and missingness will be considered later, in Sections 3.10 and 3.11.

The relationship between the manifest variables and the latent variable can be expressed as follows

$$\pi_{ijk} = \sum_t \pi^{CSLX}_{ijkt} = \pi^{CSLX}_{ijk1} + \pi^{CSLX}_{ijk2} \tag{3.52}$$

and

$$\pi^{CSLX}_{ijkt} = \pi^X_t \pi^{C|X}_{it} \pi^{S|X}_{jt} \pi^{L|X}_{kt}$$

which is given earlier as equation (3.48).

Now the hypothesis is that the system that generated the data in the 2x2x2 table of counts comes from the latent class model that satisfies equation (3.48). Since by definition,

$$\sum_t \pi^X_t = 1$$

and

$$\sum_i \pi^{C|X}_{it} = \sum_j \pi^{S|X}_{jt} = \sum_k \pi^{L|X}_{kt} = 1, \tag{3.53}$$

[3] The others can be found using the identifiability constraints, e.g. $\pi^{C|X}_{i1} = 1 - \pi^{C|X}_{i2}$ or $\lambda^{CX}_{i1} = -\lambda^{CX}_{i2}$.

it follows then that there are only seven parameters to be estimated, namely π_1^X, $\pi_{0t}^{C|X}$, $\pi_{0t}^{S|X}$ and $\pi_{0t}^{L|X}$, for $t = 1, 2$.

In keeping with the notation used in previous sections, π_{ijk} and μ_{ijk} respectively denote the proportion of individuals and the expected number of individuals in the $(i, j, k)^{th}$ cell of the 2x2x2 contingency table, so

$$\mu_{ijk} = N\pi_{ijk}.$$

It now remains to introduce the method for obtaining the maximum likelihood estimates of the parameters in the latent class model (3.48). Let $\pi_{ijkt}^{CSL|X}$ denote the conditional probability that an individual at level (i, j, k) of the manifest (observed) variables (C, S, L), will be at level t of the latent variable X. This conditional probability can be expressed in terms of the observed and latent probabilities as

$$\pi_{ijkt}^{CSL|X} = \frac{\pi_{ijkt}^{CSLX}}{\pi_{ijk}} \qquad \text{or equivalently} \qquad \pi_{ijkt}^{CSLX} = \pi_{ijk}\pi_{ijkt}^{CSL|X}. \tag{3.54}$$

From the definitions of the latent probabilities π_{ijkt}^{CSLX} and π_t^X it follows that

$$\pi_t^X = \sum_i \sum_j \sum_k \pi_{ijkt}^{CSLX}. \tag{3.55}$$

Similarly, from the definition of the conditional probabilities,

$$\pi_{it}^{C|X} = \frac{\sum_j \sum_k \pi_{ijkt}^{CSLX}}{\pi_t^X},$$

$$\pi_{jt}^{S|X} = \frac{\sum_i \sum_k \pi_{ijkt}^{CSLX}}{\pi_t^X}, \tag{3.56}$$

$$\pi_{kt}^{L|X} = \frac{\sum_i \sum_j \pi_{ijkt}^{CSLX}}{\pi_t^X}.$$

From equations (3.54), (3.55) and (3.56) supposing the conditional probabilities $\pi_{ijkt}^{CSL|X}$ are known, it becomes possible to estimate the model parameters π_t^X, $\pi_{it}^{C|X}$, $\pi_{jt}^{S|X}$ and $\pi_{kt}^{L|X}$. This is accomplished by replacing π_{ijk} with the observed probabilities p_{ijk}, which is true should the hypothesis hold. The maximum likelihood estimates of the model parameters are therefore given by the system of equations

$$\hat{\pi}_t^X = \sum_i \sum_j \sum_k p_{ijk}\hat{\pi}_{ijkt}^{CSL|X} \tag{3.57}$$

and

$$\hat{\pi}_{it}^{C|X} = \frac{\sum_j \sum_k p_{ijk}\hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X},$$

$$\hat{\pi}_{jt}^{S|X} = \frac{\sum_i \sum_k p_{ijk}\hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X}, \tag{3.58}$$

$$\hat{\pi}_{kt}^{C|X} = \frac{\sum_i \sum_j p_{ijk} \hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X}.$$

Hence, for a given set of numerical values for $\pi_{ijkt}^{CSL|X}$, tentative values of the maximum likelihood estimates of π_t^X, $\pi_{it}^{C|X}$, $\pi_{jt}^{S|X}$ and $\pi_{kt}^{L|X}$ can be obtained after inserting into (3.57) and (3.58), which in turn can be respectively inserted into (3.54), (3.55) and (3.56), to get tentative values $\hat{\pi}_{ijkt}^{CSLX}$, $\hat{\pi}_{ijk}$ and $\hat{\pi}_{ijkt}^{CSL|X}$. The initial numerical values of $\pi_{ijkt}^{CSL|X}$ can now be replaced by this new estimate, $\hat{\pi}_{ijkt}^{CSL|X}$, and then use can be made of (3.57) and (3.58) to estimate the model parameters. These new estimates can now be inserted into (3.54), (3.55) and (3.56) which would yield a new estimate of $\hat{\pi}_{ijkt}^{CSL|X}$. This process of alternating between the two sets of equations will be continued until the estimates $\hat{\pi}_t^X$, $\hat{\pi}_{it}^{C|X}$, $\hat{\pi}_{jt}^{S|X}$ and $\hat{\pi}_{kt}^{L|X}$, $\hat{\pi}_{ijkt}^{CSL|X}$, $\hat{\pi}_{ijkt}^{CSLX}$ and $\hat{\pi}_{ijk}$ remain unchanged. The estimates obtained by this iterative procedure will provide a solution to the system of equations, and if the parameters in the latent class model have maximum likelihood estimates, then Goodman (1974) proved that the maximum likelihood estimates satisfy the system of equations.

Although not explicitly put forward in Goodman (1974), his method for obtaining the maximum likelihood estimates of the parameters in the latent class model uses the EM algorithm (Dempster, Laird and Rubin (1977)). Since the missing information is the values of the latent variables, the EM algorithm functions by finding these values that maximize the joint likelihood (3.49). Therefore, starting with some initial values $\{\pi_t^{X(0)}, \pi_{it}^{C|X(0)}, \pi_{jt}^{S|X(0)}, \pi_{kt}^{L|X(0)}\}$ the procedure is to write down the joint likelihood given the observed manifest variables and the unobserved latent variables.

The log-likelihood (3.50) is replaced by its expected value (**E-step**) conditional on the observed variables. The expected values are estimated with the current values of the parameters at that iteration. Next this modified likelihood is maximized (**M-step**) to give new values for the parameters, and the whole procedure of consecutive **E** and **M-steps** is iterated until convergence. In the earlier discussion in Section 3.6 it was shown that as a result of the maximizing property of the **M-step**, even though the global maximum is not always guaranteed, the marginal likelihood over the missing (i.e. latent) variable will never decrease with each iteration.

Consequently the **E-step** is made up of three sub-steps

$$\hat{\pi}_{ijkt}^{CSLX} = \hat{\pi}_t^X \hat{\pi}_{it}^{C|X} \hat{\pi}_{jt}^{S|X} \hat{\pi}_{kt}^{L|X}, \qquad \text{(E1)}$$

$$\hat{\pi}_{ijk} = \sum_t \hat{\pi}_{ijkt}^{CSLX}, \qquad \text{(E2)}$$

$$\hat{\pi}_{ijkt}^{CSL|X} = \frac{\hat{\pi}_{ijkt}^{CSLX}}{\hat{\pi}_{ijk}}. \qquad \text{(E3)}$$

The observed cell probabilities p_{ijk} can be used to obtain new trial values in the **M-step** comprising of four sub-steps

$$\hat{\pi}_t^X = \sum_i \sum_j \sum_k p_{ijk} \hat{\pi}_{ijkt}^{CSL|X}, \qquad \text{(M1)}$$

$$\hat{\pi}_{it}^{C|X} = \frac{\sum_j \sum_k p_{ijk} \hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X}, \qquad \text{(M2)}$$

$$\hat{\pi}_{jt}^{S|X} = \frac{\sum_i \sum_k p_{ijk} \hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X}, \qquad \text{(M3)}$$

$$\hat{\pi}_{kt}^{L|X} = \frac{\sum_i \sum_j p_{ijk} \hat{\pi}_{ijkt}^{CSL|X}}{\hat{\pi}_t^X}. \qquad \text{(M4)}$$

Note that the iterative proportional fitting algorithm (IPF) can alternatively be used to obtain maximum likelihood estimates.

3.9.2 Maximum Likelihood Estimation of Parameters under the Haberman Parameterization for the Local Independence Latent Model

Denoting the parameters of the model by η and τ, the log-linear representation of the latent class model is

$$\mu_{ijkt}^{CSLX} = \eta \tau_i^C \tau_j^S \tau_k^L \tau_t^X \tau_{it}^{CX} \tau_{jt}^{SX} \tau_{kt}^{LX} \tag{3.59}$$

or equivalently

$$\log \mu_{ijkt}^{CSLX} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)}.$$

This is not an identifiable model; there are more unknown parameters to be estimated than known cell frequencies in the 2x2x2 contingency table. Identifying restrictions on the parameters are necessary, and can be accomplished by constraining the sums of each λ-parameter to zero. These sum-to-zero constrains are represented as follows:

$$\sum_i \lambda_i^{(C)} = \sum_j \lambda_j^{(S)} = \sum_k \lambda_k^{(L)} = \sum_t \lambda_t^{(X)} = 0$$

and

$$\sum_i \lambda_{it}^{(CX)} = \sum_j \lambda_{jt}^{(SX)} = \sum_k \lambda_{kt}^{(LX)} = \sum_t \lambda_{it}^{(CX)} = \sum_t \lambda_{jt}^{(SX)} = \sum_t \lambda_{kt}^{(LX)} = 0.$$

With the identifying restrictions imposed on the parameters, an exactly identified model results that requires no additional restrictions on the data. Under this model the observed cell frequencies are the same as the maximum likelihood estimates of the expected cell frequencies. Therefore, given the observed 2x2x2 table of counts n_{ijk} the objective is to find the 2x2x2x2 latent contingency table with cell counts n_{ijkt} and μ_{ijkt}^{CSLX} that satisfy (3.59), subject to the sum-to-zero constraints. For this model the maximum likelihood estimates are given by

$$\hat{\mu}_{i++t}^{CSLX} = n_{i++t}, \quad \hat{\mu}_{+j+t}^{CSLX} = n_{+j+t} \quad \text{and} \quad \hat{\mu}_{++kt}^{CSLX} = n_{++kt} \tag{3.60}$$

where the marginal observed sums are given by

$$n_{i++t} = \sum_j \sum_k n_{ijkt}, \quad n_{+j+t} = \sum_i \sum_k n_{ijkt} \quad \text{and} \quad n_{++kt} = \sum_i \sum_j n_{ijkt}.$$

The marginal expected sums μ_{i++t}^{CSLX}, μ_{+j+t}^{CSLX} and μ_{++kt}^{CSLX} are found in a similar manner. However, n_{i++t}, n_{+j+t} and n_{++kt} are not observed but when given n_{ijk}, the expected values of latent table counts n_{ijkt} are

$$\hat{n}_{ijkt} = \frac{n_{ijk}}{\hat{\mu}_{ijk}} \hat{\mu}_{ijkt}.$$

Thus, the maximum likelihood equations can now be written as

$$\hat{\mu}_{i++t}^{CSLX} = \hat{n}_{i++t}, \quad \hat{\mu}_{+j+t}^{CSLX} = \hat{n}_{+j+t} \quad \text{and} \quad \hat{\mu}_{++kt}^{CSLX} = \hat{n}_{++kt}. \tag{3.61}$$

The iterative proportional fitting algorithm (IPF) is used to find the maximum likelihood estimates of the cell counts such that (3.59) holds. As per usual the IPF algorithm works by iteratively adapting the initial estimates of the expected cell frequencies to the observed marginal frequencies to be reproduced by the given model. What is different here is that the marginal frequencies are 'known' as a result of the latent variable. Additionally, the starting values for the iterative computations play a greater role because of the model constraints imposed; without a correct choice of starting values the iterative process proceeds very slowly and may not converge to the maximum likelihood estimates. Hence to begin the algorithm some initial estimates are required that satisfy (3.59) and the sum-to-zero constraints such that

$$\log \mu_{ijkt}^{CSLX(0)} = \lambda^{(0)} + \lambda_i^{C(0)} + \lambda_j^{S(0)} + \lambda_k^{L(0)} + \lambda_t^{X(0)} + \lambda_{it}^{CX(0)} + \lambda_{jt}^{SX(0)} + \lambda_{kt}^{LX(0)}. \tag{3.62}$$

For notational simplicity, $\mu_{ijkt}^{CSLX(r)}$ is replaced with $\mu_{ijkt}^{(r)}$ where r represents the rth iteration. Clearly the initial estimates need to be chosen such that they are in agreement with the hypothesized model (3.59), so given the observed cell counts n_{ijk} it becomes possible to find initial estimates $\mu_{ijkt}^{(0)}$ that satisfy the sum-to-zero constraints, and then subsequently find the $n_{ijkt}^{(0)}$. Given these initial estimates $\mu_{ijkt}^{(0)}$ and $n_{ijkt}^{(0)}$, define $n_{i++t}^{(r)}$, $n_{+j+t}^{(r)}$ and $n_{++kt}^{(r)}$ as the estimated marginal observed counts at the rth iteration; and $\mu_{i++t}^{(r)}$, $\mu_{+j+t}^{(r)}$ and $\mu_{++kt}^{(r)}$ as the respective estimated marginal expected counts. The cycle is initialized by

$$n_{ijkt}^{(0)} = \mu_{ijkt}^{(0)} \times \frac{n_{ijk}}{\mu_{ijk}^{(0)}}, \tag{1}$$

$$\mu_{ijkt}^{(1)} = \mu_{ijkt}^{(0)} \times \frac{n_{i++t}^{(0)}}{\mu_{i++t}^{(0)}}, \tag{2}$$

$$\mu_{ijkt}^{(2)} = \mu_{ijkt}^{(1)} \times \frac{n_{+j+t}^{(0)}}{\mu_{+j+t}^{(1)}}, \tag{3}$$

$$\mu_{ijkt}^{(3)} = \mu_{ijkt}^{(2)} \times \frac{n_{++kt}^{(0)}}{\mu_{i++kt}^{(2)}}. \tag{4}$$

A new cycle begins by re-estimating the estimates such that

$$n_{ijkt}^{(1)} = \mu_{ijkt}^{(3)} \times \frac{n_{ijk}}{\mu_{ijk}^{(3)}}, \tag{5}$$

$$\mu_{ijkt}^{(4)} = \mu_{ijkt}^{(3)} \times \frac{n_{i++t}^{(1)}}{\mu_{i++t}^{(3)}}, \tag{6}$$

$$\mu_{ijkt}^{(5)} = \mu_{ijkt}^{(4)} \times \frac{n_{+j+t}^{(1)}}{\mu_{+j+t}^{(4)}}, \tag{7}$$

$$\mu_{ijkt}^{(6)} = \mu_{ijkt}^{(5)} \times \frac{n_{++kt}^{(1)}}{\mu_{++kt}^{(5)}}. \tag{8}$$

As the cycle progresses the estimated latent frequencies should come simultaneously closer such that the log-linear model is satisfied, under the restrictions, eventually yielding the maximum likelihood estimates. The reason why the initial estimates are crucial is because the IPF works by first initializing $\mu_{ijkt}^{(0)}$ where these initial estimates satisfy the log-linear model and conditions. Therefore, if these initial approximations are not correctly specified there are problems encountered during convergence of the algorithm. Generally, convergence of the IPF algorithm is much slower than with directly observed frequency counts (Haberman (1979)).

For the IPF algorithm in this application, the marginal totals, although effectively unknown, are treated as known and so the iterative procedure described above can be re-written as an EM algorithm. Here provided some initial values $\hat{\mu}_{ijkt}^{(0)}$ and $\hat{\mu}_{ijk}^{(0)}$ have been obtained, there is a single **E-step**

$$\hat{n}_{ijkt}^{(0)} = \hat{\mu}_{ijkt}^{(0)} \times \frac{n_{ijk}}{\hat{\mu}_{ijk}^{(0)}} \tag{E}$$

and the M-step is comprised of three sub-steps

$$\hat{\mu}_{ijkt}^{(1)} = \hat{\mu}_{ijkt}^{(0)} \times \frac{\hat{n}_{i++t}^{(0)}}{\hat{\mu}_{i++t}^{(0)}}, \tag{M1}$$

$$\hat{\mu}_{ijkt}^{(2)} = \hat{\mu}_{ijkt}^{(1)} \times \frac{\hat{n}_{+j+t}^{(0)}}{\hat{\mu}_{+j+t}^{(1)}}, \tag{M2}$$

$$\hat{\mu}_{ijkt}^{(3)} = \hat{\mu}_{ijkt}^{(2)} \times \frac{\hat{n}_{++kt}^{(0)}}{\hat{\mu}_{++kt}^{(2)}}. \tag{M3}$$

Haberman (1979) went on to show that the series of **M-steps** can be written as a single log-linear modelling step as

$$\log \hat{\mu}_{ijkt} = \hat{\lambda} + \hat{\lambda}_i^{(C)} + \hat{\lambda}_j^{(S)} + \hat{\lambda}_k^{(L)} + \hat{\lambda}_t^{(X)} + \hat{\lambda}_{it}^{(CX)} + \hat{\lambda}_{jt}^{(SX)} + \hat{\lambda}_{kt}^{(LX)},$$

where the $\hat{\lambda}$-terms are subject to the same identifying constraints

$$\sum_i \hat{\lambda}_i^{(C)} = \sum_j \hat{\lambda}_j^{(S)} = \sum_k \hat{\lambda}_k^{(L)} = \sum_t \hat{\lambda}_t^{(X)} = 0$$

and

$$\sum_i \hat{\lambda}_{it}^{(CX)} = \sum_j \hat{\lambda}_{jt}^{(SX)} = \sum_k \hat{\lambda}_{kt}^{(LX)} = \sum_t \hat{\lambda}_{it}^{(CX)} = \sum_t \hat{\lambda}_{jt}^{(SX)} = \sum_t \hat{\lambda}_{kt}^{(LX)} = 0.$$

3.10 Some Issues with the Latent Class Framework in Triple System Estimation

Latent class models rely on the basic assumption of local independence between the latent and manifest variables. Obviously in the census scenario with bias introduced due to the lack of independence between the Census and Survey, the local independence latent class model is not ideal. Here, a new model specification is undertaken when local independence is clearly violated. Under the conventional latent class model, the latent variable explains all the association between the manifest variables. In a local dependence model there is some residual unexplained association. In the initial work on latent class analysis by Lazarsfeld and Henry (1968), Goodman (1974) and Haberman (1979) this assumption was essential in the derivation of the latent class model. However, the constraints imposed in order to have local independence may be unrealistic (Hagenaars (1993)). In the three-sample capture-recapture model, there may be some residual association between some of the lists, even after taking the latent variable into account. For example there could be a pairwise association between the Census and Survey, leading to the contingency tables shown in Tables 3.9 and 3.10.

The motivation for this latent model stems from the fact that some dependence may be introduced between the Census and Survey, firstly, due to the reactions of individuals to the census enumeration process and secondly, due to associations in the census and survey logistical operations. Enumeration in the Census or Survey depends largely on an individual's attitude to being interviewed (or filling in a form) and their social responsibility in general. The listing of an individual on an administrative records list, on the other hand, usually depends on factors that provide direct benefit to the individual. This is certainly true for administrative records that provide tax rebates and unemployment allowances. Although there may be an issue about the timing, this does hold true to a certain extent in health registers; a person might forget to contact their general practitioner to update their medical records but will do so when they are ill. It does, therefore, follow that whether a person appears on the administrative list should not be greatly influenced by the individual's choice or ability to participate in the Census or Survey[4].

Table 3.9: Local dependence latent class model

		Class 1 - Real		Class 2 - Erroneous	
		Third List			
		Counted	Missed	Counted	Missed
	Counted in both	n_{1111}	n_{1101}	n_{1112}	n_{1102}
Census and Survey	Counted in Census, Missed in Survey	n_{1011}	n_{1001}	n_{1012}	n_{1002}
	Missed in Census, Counted in Survey	n_{0111}	n_{0101}	n_{0112}	n_{0102}
	Missed in both	n_{0011}	n_{0001}	n_{0012}	n_{0002}

[4] Admittedly, this does break down when dealing with people who are perhaps considered to exist on the fringes of society, i.e. those not registered on any administrative lists and also fail to participate in the Census or Survey processes.

Table 3.10: Conditional Parameterization with local dependence

	Survey				Third List	
	Counted		Missed			
	Census		Census		Counted	Missed
	Counted	Missed	Counted	Missed		
Class 1	$\pi_{111}^{CS\|X}$	$\pi_{011}^{CS\|X}$	$\pi_{101}^{CS\|X}$	$\pi_{001}^{CS\|X}$	$\pi_{11}^{L\|X}$	$\pi_{01}^{L\|X}$
Class 2	$\pi_{112}^{CS\|X}$	$\pi_{012}^{CS\|X}$	$\pi_{102}^{CS\|X}$	$\pi_{002}^{CS\|X}$	$\pi_{12}^{L\|X}$	$\pi_{02}^{L\|X}$

The local independence model given in (3.48) becomes

$$\begin{aligned} \pi_{ijkt}^{CSLX} &= \pi_t^X \pi_{ijt}^{CS|X} \pi_{kt}^{L|X} \\ \implies \quad \pi_{ijk} &= \sum_t \pi_t^X \pi_{ijt}^{CS|X} \pi_{kt}^{L|X} \end{aligned} \tag{3.63}$$

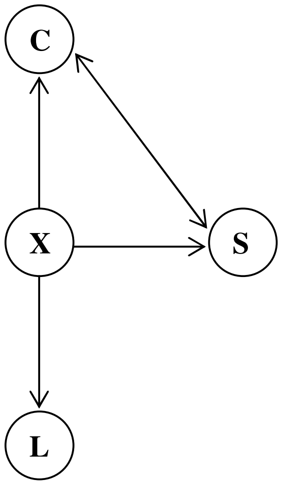
where $\pi_{ijt}^{CS|X}$ is the conditional probability of being at level i of the Census variable and level j of the Survey and is given by

$$\pi_{ijt}^{CS|X} = \Pr\left(C = i, S = j | X = t\right) = \frac{\exp\left(\lambda_{ij}^{(CS)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(LX)}\right)}{\sum_i \sum_j \exp\left(\lambda_{ij}^{(CS)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(LX)}\right)}.$$

So individuals can still be classified into mutually exclusive and exhaustive latent classes, but within each class the independence assumption between the indicators needs to be relaxed to include some additional dependence between the Census and Survey. This is accomplished by allowing the latent class model to incorporate an additional direct effect that can account for any additional association between C and S that is not explainable under (3.48); as shown by Figure 3.5. Therefore, the local dependence model is a standard latent class model with one dichotomous latent variable X, and two manifest variables CS and L. Here, within each latent class, CS and L are independent of each other, but the difference is that elements of C and S of the joint variable CS are permitted to be associated within latent classes. Therefore, the resulting manifest variables CS and L are correlated with each other but this correlation disappears when the latent variable is taken in account.

Hagenaars (1993) suggests accounting for this unexplained variation in terms of an additional latent variable. However, this is rejected in the triple system scenario as an additional latent variable further exacerbates the issue of model identifiability. Another way, which will be considered here, is to observe that the remaining variation is mostly due to some additional association in the Census and Survey that the latent variable fails to account for, so by adding a direct effect between the Census and Survey corrects for this. The log-linear formulation of the latent class model makes it easy to conceptualize how the direct effects between the manifest variables can be incorporated. All that is needed is to introduce the extra effect parameters that represent the desired direct effects among the manifest variables into the log-linear model.

Figure 3.5: Independence graph showing the relationship between the latent and manifest variables - with dependence between the Census and Survey



Thus the local dependence model (3.63) can be written as the log-linear model

$$\log \mu_{ijkt} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)} + \lambda_{ijt}^{(CSX)}, \quad (3.64)$$

with constraints

$$\sum_i \lambda_i^{(C)} = \sum_j \lambda_j^{(S)} = \sum_k \lambda_k^{(L)} = \sum_t \lambda_t^{(X)} = 0$$

and

$$\sum_i \lambda_{it}^{(CX)} = \sum_j \lambda_{jt}^{(SX)} = \sum_k \lambda_{kt}^{(LX)} = \sum_i \lambda_{ij}^{(CS)} = \sum_j \lambda_{ij}^{(CS)} = 0,$$

$$\sum_t \lambda_{it}^{(CX)} = \sum_t \lambda_{jt}^{(SX)} = \sum_t \lambda_{kt}^{(LX)} = 0$$

and

$$\sum_i \sum_j \lambda_{ijt}^{(CSX)} = \sum_i \sum_t \lambda_{ijt}^{(CSX)} = \sum_j \sum_t \lambda_{ijt}^{(CSX)} = 0.$$

The interpretation of the local dependence model, written in the form (3.64) implies that the similarity among responses is caused by some subject-specific factors, operating together with the latent category, and a failure to account for this could lead to a wrong model being specified. Unfortunately, the above log-linear model (3.64) is not identified as there are too many unknowns so some conditions are required in order that the parameters can be estimated. The way forward is to apply some identification conditions to the model. If the CSX-interaction effect is significantly small (3.64) could be represented by $\{LX, CX, SX, CS\}$. So the ensuing log-linear model becomes

$$\log \mu_{ijkt} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)}. \quad (3.65)$$

However, even with this slight modification the local dependence model given in (3.65) is still not identifiable - it has nine unknowns and eight observed cells[5]. There are a number of restrictions that can be placed on the parameters to ensure that the model becomes identifiable. Goodman (1974) and Hagenaars (1993) give a wide range of these restrictions. For example, it may be assumed that the latent classes are equiprobable, and the process that drives whether a person is a real or erroneous enumeration is the same. Under this latent class model, the latent classes are completely specified, so

$$\pi_t^X = \frac{1}{2} \quad \text{or equivalently} \quad \exp\left(\lambda + \lambda_t^{(X)}\right) = \frac{1}{2}. \qquad (R1)$$

Another restriction is to set

$$\begin{aligned} &\pi_{01}^{CX} = \pi_{12}^{CX}, \quad \pi_{01}^{SX} = \pi_{12}^{SX} \quad \text{and} \quad \pi_{01}^{LX} = \pi_{12}^{LX} \\ &\text{or equivalently} \qquad\qquad\qquad\qquad\qquad\qquad (R2) \\ &\lambda_1^{(C)} = \lambda_1^{(S)} = \lambda_1^{(L)} = 1. \end{aligned}$$

The interpretation of this restriction is that it is believed that the conditional probabilities of being missed given that an individual is real is the same as the probabilities of being counted given that the individual is erroneous. In other words, the conditional probability of being counted given person is a real enumeration is equal to that of a person being missed given that they are an erroneous enumeration. This restriction may be too stringent as it fails to acknowledge known differences between the Census, Survey and Third List enumeration processes. A lesser, more realistic, restriction could be to constrain only the Survey conditional probabilities. Here

$$\pi_{01}^{SX} = \pi_{12}^{SX} = 0 \quad \text{or equivalently} \quad \lambda_1^{(S)} = 1. \qquad (R3)$$

Albeit these restrictions reduce the number of parameters that need to be estimated since some of the parameters are specified, they do not all manage to achieve model identifiability - only $R2$ leaves model (3.65) identified, since there are six estimable parameters, and eight cells[6]. However, $R3$ is the most intuitively appealing in a triple system census application. This is because in an adequately designed post-enumeration survey it is not unfeasible to assume that all the Survey enumerations are error-free; effectively, all the individuals counted by the Survey are real enumerations. Unfortunately, this local dependence model with the $R3$ restriction is still not fully identifiable.

[5] This identifiability problem is further compounded when dealing with the incomplete 2x2x2 contingency table, where there are only seven observable.

[6] It must be borne in mind that in actuality there are only seven observable cells in the incomplete 2x2x2 contingency table.

3.11 Proposed Solution to Local Dependence Model Non Identifiability

The previous discussion has shown that the major problem encountered in the use of the latent model in specifying whether an individual is an erroneous or real enumeration concerns model identifiability. The local independence model for the 2x2x2 (with the n_{000} cell assumed to be observed) is exactly identified, with zero degrees of freedom. But, it has been argued that the most suitable model to be fitted when there are three systems - the Census, Survey and Third List - is the local dependence model with an additional pairwise association term between the Census and Survey. It was also shown earlier, in Section 3.10, that this model is non-identified as there are too many parameters. Although it is possible to place restrictions on some of the parameters to ensure identifiability, the realization is that most of these restrictions are not intuitively appealing. The issue here is simply that there are too few cell counts for the required model.

Table 3.11: Contingency table of the local dependence latent class model with the gender covariate

Males

		Class 1		Class 2	
		Third List		Third List	
		Counted	Missed	Counted	Missed
	Counted in both	n_{1111m}	n_{1101m}	n_{1112m}	n_{1102m}
Census and Survey	Counted in Census, Missed in Survey	n_{1011m}	n_{1001m}	n_{1012m}	n_{1002m}
	Missed in Census, Counted in Survey	n_{0111m}	n_{0101m}	n_{0112m}	n_{0102m}
	Missed in both	n_{0011m}	n_{0001m}	n_{0012m}	n_{0002m}

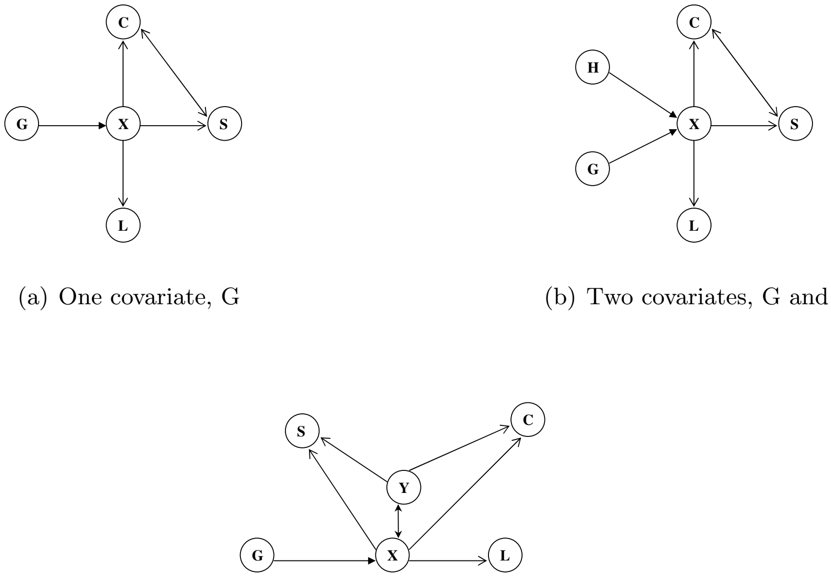
Females

		Class 1		Class 2	
		Third List		Third List	
		Counted	Missed	Counted	Missed
	Counted in both	n_{1111f}	n_{1101f}	n_{1112f}	n_{1102f}
Census and Survey	Counted in Census, Missed in Survey	n_{1011f}	n_{1001f}	n_{1012f}	n_{1002f}
	Missed in Census, Counted in Survey	n_{0111f}	n_{0101f}	n_{0112f}	n_{0102f}
	Missed in both	n_{0011f}	n_{0001f}	n_{0012f}	n_{0002f}

In the 2001 Census, post-stratification by age, gender and other covariates ensured that whilst performing the dual system estimation all individuals in the 2x2 contingency table could be assumed to have the same capture probabilities. Instead of post-stratification, the log-linear modelling framework makes it possible to directly include the covariates that introduce heterogeneity into the capture probabilities. So as a representation, consider the triple system case with a latent variable, then the contingency table stratified by gender is as shown in Table 3.11.

This can be interpreted that once a person's gender been accounted for, the relationships amongst the Census, Survey and Third List with the latent variable is the same in both male and female sub-tables. This additional information has the effect of freeing up some degrees of freedom, but this covariate can be difficult to conceptualize since it is only supposed to be related to the latent variable, that is the effect of gender on the Census, Survey and Third List is completely mediated through the latent variable, as shown in Figure 3.6, below.

Figure 3.6: Simultaneous Latent Class Models



(a) One covariate, G

(b) Two covariates, G and H

(c) Two latent variables, X and Y

Evidently, Figure 3.6(a) can be written under the Haberman parameterization as

$$\log \mu_{ijktg}^{(CSLXG)} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{ij}^{(CS)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)}. \quad (3.66)$$

It is nonetheless difficult to write the model represented under the Goodman parameterization; this is one advantage of using the Haberman log-linear parameterization of latent class models. It follows that if further interaction terms are needed and there is a lack of degrees of freedom, then other covariates (as shown in Figure 3.6(b)) can be introduced into the analysis, but it still remains that these covariates need to be related only to the latent variable, and not the manifest variables. So similarly, Figure 3.6(b) can be written as the log-linear model

$$\log \mu_{ijktgh}^{(CSLXGH)} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_h^{(H)} + \lambda_{ij}^{(CS)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)} + \lambda_{ht}^{(HX)}. \quad (3.67)$$

This is what McCutcheon (1987) refers to as the *simultaneous latent class model.* Clearly, this could introduce additional complexities in the interpretation, and McCutcheon (1987) and Hagenaars (1993) suggest framing the problem as a two-latent variable problem. This model, shown in Figure 3.6(c) has an additional latent variable, Y, brought in to account for the association between the Census and Survey not fully accounted for by the first latent variable, X. This modification is a very interesting alternative model as it implies that a substantive interpretation can now be sought for this ‘new’ latent variable, for instance it could represent an individual's propensity to participate in the Census or Survey, which is expected to be different from their participation in an Administrative List due to the different enumeration mechanisms involved. However, this model, written down in equation (3.68), is not a viable solution in the problem under study due to identifiability issues.

$$\log \mu^{(CSLXYG)}_{ijktyg} = \lambda + \lambda^{(C)}_{i} + \lambda^{(S)}_{j} + \lambda^{(L)}_{k} + \lambda^{(X)}_{t} + \lambda^{(Y)}_{y} + \lambda^{(G)}_{g} + \lambda^{(CX)}_{it} + \lambda^{(SX)}_{jt} + \lambda^{(LX)}_{kt} + \lambda^{(GX)}_{gt} + + \lambda^{(CY)}_{iy} + \lambda^{(SY)}_{jy}. \tag{3.68}$$

3.12 A Note on Bayesian Methods for Capture-Recapture Models

Oftentimes there is some historical information available regarding the population size N - for example the One Number Census estimated the population of Scotland at 5,062,011, and it is not completely unreasonable that this is used as a prior estimate in the determination of the Scottish population at the next census. This goes to show that the capture-recapture model intuitively lends itself to be formulated in the Bayesian paradigm, seeing as information about the total population size can be regarded as being updated from one sampling occasion to the next. Bayesian statistics does, therefore, provide a mathematical framework for revising knowledge and Smith (1988) demonstrated that it is possible to find Bayesian estimators of the population size that resemble the traditional Lincoln-Petersen estimators. Moreover, estimates of precision can be computed in the Bayesian paradigm without the assumption of normality, which is often needed in the classical capture-recapture case.

To formulate the basic two sample capture-recapture model under the Bayesian framework, first think of N as being constant, and each individual as having an equi-probable chance of being in either the first or second sample. (These are basically the same assumptions required under the classical framework.) The conditional distribution of being in both the first and second samples, given the first and second sample observed totals, can be written as the hypergeometric distribution

$$f\left(n_{11}|n_{1+}, n_{+1}\right) = \frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix} \begin{pmatrix} N - n_{1+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} N \\ n_{+1} \end{pmatrix}}. \tag{3.69}$$

Seber (1982) showed that under certain conditions (i.e. that the samples are large and the number of individuals found in both samples as a proportion of the population is small[7]),

[7] This condition will not be entirely appropriate with regards to a human census. For example, in the 2001 UK Census even in areas where the census was deemed to have performed badly, the initial census enumeration achieved over 60% coverage (Office for National Statistics (2004)).

this likelihood can be satisfactorily approximated by the Poisson density

$$f\left(n_{11}|\frac{1}{N},n_{1+},n_{+1}\right)=\frac{\left(\frac{n_{1+}n_{+1}}{N}\right)^{n_{11}}\exp\left(-\frac{n_{1+}n_{+1}}{N}\right)}{n_{11}!}. \tag{3.70}$$

Since n_{11}, n_{+1} and n_{1+} are observed counts but N is unknown, any prior information about the population size may now be incorporated using a Gamma density,

$$f\left(\frac{1}{N}\right)=\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\left(\frac{1}{N}\right)^{\alpha-1}\exp\left(-\frac{\beta}{N}\right) \tag{3.71}$$

which is a conjugate of the Poisson density. It follows that the posterior is given by

$$\begin{aligned} f\left(\frac{1}{N}|n_{11},n_{+1},n_{1+}\right) &\propto \frac{\left(\frac{n_{+1}n_{1+}}{N}\right)^{n_{11}}\exp\left(-\frac{n_{+1}n_{1+}}{N}\right)}{n_{11}!}\times\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\left(\frac{1}{N}\right)^{\alpha-1}\exp\left(-\frac{\beta}{N}\right) \\ &\propto \left(\frac{1}{N}\right)^{\alpha+n_{11}-1}\exp\left(-\frac{\beta+n_{+1}n_{1+}}{N}\right) \end{aligned}$$

which is a Gamma(α^*, β^*), where $\alpha^*=\alpha+n_{11}$ and $\beta^*=\beta+n_{+1}n_{1+}$.

The posterior mean of the Gamma distribution is

$$\frac{\alpha^*}{\beta^*}=\frac{\alpha+n_{11}}{\beta+n_{+1}n_{1+}}=\left(\frac{\beta}{\beta+n_{+1}n_{1+}}\right)\times\frac{\alpha}{\beta}+\left(1-\frac{\beta}{\beta+n_{+1}n_{1+}}\right)\times\frac{n_{11}}{n_{+1}n_{1+}}. \tag{3.72}$$

Equation (3.72) is a weighted average of the prior mean and the classical estimator, of form $w\times$ prior mean + $(1-w)\times$ m.l.e.; on closer inspection this sample mean (equivalent to the maximum likelihood estimate, here) is the same as the reciprocal of the Lincoln-Petersen estimator of the population size. Further, since $n_{+1}n_{1+}$ is generally quite large, the data does dominate the posterior, but the advantage the Bayesian paradigm has over its classical counterpart is that suitable choices of α and β can be selected to reflect the extent of the prior knowledge concerning N. Therefore, the posterior distribution effectually incorporates the information from the data and any subjective prior knowledge.

This can be generalised for the case where there are more than two captures. Firstly, define r to be the sampling occasion. If on the rth sampling occasion n_r individuals are captured and m_r represents the number of individuals that have previously been captured. Define M_r to be the total number of marked individuals in the population, just before sample r is taken. Then the conditional distribution of the number of marked individuals is a hypergeometric in the sample r, given by

$$f\left(m_r|M_r,N\right)=\frac{\begin{pmatrix}M_r\\m_r\end{pmatrix}\begin{pmatrix}N-M_r\\n_r-m_r\end{pmatrix}}{\begin{pmatrix}N\\n_r\end{pmatrix}}. \tag{3.73}$$

Equation (3.73) is basically the same as that for the two-sample case, and so for multiple captures the product hypergeometric results

$$f\left(m_r|M_r,N\right)=\prod_r\frac{\begin{pmatrix}M_r\\m_r\end{pmatrix}\begin{pmatrix}N-M_r\\n_r-m_r\end{pmatrix}}{\begin{pmatrix}N\\n_r\end{pmatrix}} \tag{3.74}$$

which can be approximated as the Poisson density

$$f\left(m_r|n_r, M_r, \frac{1}{N}\right) = \frac{\left(\frac{\sum_r n_r M_r}{N}\right)^{\sum_r m_r} \exp\left(-\frac{\sum_r n_r M_r}{N}\right)}{(\sum_r m_r)!}. \tag{3.75}$$

After choosing a Gamma prior, similar to (3.71) and using the fact that the Poisson and Gamma are conjugates, the posterior of the multiple capture sample size is

$$f\left(\frac{1}{N}|m_r, n_r, \sum_r M_r\right) \propto \left(\frac{1}{N}\right)^{\alpha-1+\sum_r m_r} \exp\left\{-\frac{1}{N}\left(\beta + \sum_r n_r M_r\right)\right\} \tag{3.76}$$

with posterior mean given by

$$\frac{\alpha + \sum_r m_r}{\beta + \sum_r n_r M_r} = \left(\frac{\beta}{\beta + \sum_r n_r M_r}\right)\frac{\alpha}{\beta} + \left(1 - \frac{\beta}{\beta + \sum_r n_r M_r}\right)\frac{\sum_r m_r}{\sum_r n_r M_r}. \tag{3.77}$$

This is a weighted average of the prior mean and the reciprocal of the Schnabel estimator (Schnabel (1938)).

The posteriors of the population size estimated using equations (3.77) and (3.72) under the Bayesian paradigm both assume list independence. However, Nandram and Zelterman (2007) and King and Brooks (2001) use different (more computationally intensive) techniques to calculate the posterior distribution when heterogeneity needs to be introduced due to the dependence between the samples. Nandram and Zelterman (2007) place prior distributions on the marginal probabilities of being present on the lists and the odds representing the marginal interactions, and uses rejection sampling to find the target posterior distribution. King and Brooks (2001) on the other hand fit a Bayesian log-linear model to the capture-recapture data. Here instead of specifying priors on the cell counts, priors are placed on the log-linear parameters. Then using reversible jump Markov Chain Monte Carlo (MCMC) techniques the best model is chosen by model averaging. Bayesian model averaging accounts for the uncertainty inherent in the model selection process by averaging over many different, often competing, models (Dellaportas and Forster (1999)). It therefore incorporates model uncertainty into the conclusions about parameters and prediction.

The ideas behind the rejection sampling can now be illustrated in a triple system context. For an incomplete 2x2x2 contingency table with cell probabilities and cell counts represented by $\{\pi_{000}, \pi_{001}, \pi_{010}, \pi_{011}, \pi_{100}, \pi_{101}, \pi_{110}, \pi_{111}\}$ and $\{n_{000}, n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}\}$, with π_{000} and n_{000} representing the unobserved cell, as per usual. Now let $\{\pi'_{001}, \pi'_{010}, \pi'_{011}, \pi'_{100}, \pi'_{101}, \pi'_{110}, \pi'_{111}\}$ be the probabilities of the seven observable cells, where it can be recalled that $\pi'_{ijk} = \frac{\pi_{ijk}}{1-\pi_{000}}$. Finally define θ to be the three marginal cell probabilities, η_i^C, η_j^S, η_k^L, the three marginal pairwise association terms ψ_{ij}^{CS}, ψ_{ik}^{CL}, ψ_{jk}^{SL} and the three-way association term ψ_{ijk}^{CSL}. Then the likelihood of the observed cell counts given θ is

$$\begin{aligned}\ell\left(n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}|\eta_i^C, \eta_j^S, \eta_k^L, \psi_{ij}^{CS}, \psi_{ik}^{CL}, \psi_{jk}^{SL}, \psi_{ijk}^{CSL}\right) &= \prod_S \left(\pi'_{ijk}\right)^{n_{ijk}} \\ &= \prod_S \left(\frac{\pi_{ijk}}{1-\pi_{000}}\right)^{n_{ijk}} \end{aligned} \tag{3.78}$$

where π_{ijk} are functions of θ and S is the set of all cells apart from the (0,0,0) cell.

Now the Census marginal probability is given by

$$\eta_i^C = \pi_{1++} = \sum_j \sum_k \pi_{1jk},$$

the Survey marginal probability is

$$\eta_j^S = \pi_{+1+} = \sum_i \sum_k \pi_{i1k},$$

and the List marginal probability is

$$\eta_k^L = \pi_{++1} = \sum_i \sum_j \pi_{ij1}.$$

Define the marginal odds ratios to be

$$\psi_{ij}^{CS} = \frac{\pi_{11+}\pi_{00+}}{\pi_{10+}\pi_{01+}}, \qquad \psi_{ik}^{CL} = \frac{\pi_{1+1}\pi_{0+0}}{\pi_{1+0}\pi_{0+1}}, \qquad \psi_{jk}^{SL} = \frac{\pi_{+11}\pi_{+00}}{\pi_{+10}\pi_{+01}},$$

and

$$\psi_{ijk}^{CSL} = \frac{\pi_{111}\pi_{100}\pi_{010}\pi_{001}}{\pi_{110}\pi_{101}\pi_{011}\pi_{000}}.$$

It is easy to show that (3.78) is maximized at

$$\ell_{max} = \prod_S \left(\frac{n_{ijk}}{n}\right)^{n_{ijk}}, \qquad \text{where } n \text{ is the sum of the observed cell counts.} \qquad (3.79)$$

Since there is a missing cell, in the classical framework it is assumed that there is no three-way interaction term. This assumption can be relaxed under the Bayesian framework. Here, the rejection sampler is used, but there are a number of other ways the posterior distribution of N can be found - e.g. the Metropolis Hastings algorithm, Gibbs Sampler or other Markov Chain Monte Carlo (MCMC) methods. The rejection sampler used here generates θ, i.e. seven numbers from the priors representing the marginal cell probabilities and the marginal odds ratios. Given this set of numbers, the iterative proportional fitting algorithm is used to find eight cell probabilities $\{p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}\}$ that satisfy θ. At the next stage of the rejection sampler, this generated sample is accepted with probability $\frac{\ell}{\ell_{max}}$ as an observation from the posterior distribution. Otherwise this sample is rejected and another sample is generated from the seven priors. This process is repeated until the number of samples is large enough for the posterior distribution to be accurately estimated. When implementing this, it was found that there are difficulties surrounding the acceptance and the algorithm takes a long time before producing a posterior, as a large number of the generated contingency tables are rejected. This is further exacerbated by the fact that a great deal of effort can be devoted to the construction and choice of a distribution that characterizes the available prior information without yielding any substantial benefits (Nandram and Zelterman (2007)).

The above has not considered the case where there is overenumeration. Nonetheless, since the previous section discussed how to formulate the contingency table with overenumeration as a latent class model, which can be represented as a log-linear model it does follow that the Bayesian log-linear modelling techniques discussed by King and Brooks (2001) and Dellaportas and Forster (1999), amongst others, can be extended to the latent class models.

The main idea behind the Bayesian approach is to construct a joint prior distribution over the unknown quantities, and so one advantage the Bayesian paradigm has is that it can cope with non-identifiable models. The earliest application of Bayesian models to latent class data was a paper by Evans et al. (1989) who fitted a latent class model to a 2x2 contingency table. Traditionally, latent class analysis has been concentrated to the case where there are at least three manifest variables, because of the issues surrounding model identifiability (Goodman (1974)). Latent models with two manifest variables have been predominantly discounted because they rely on a number of assumptions to ensure identifiability. What Evans et al. (1989) do is to place (both informative and non-informative) priors on the latent class and conditional response probabilities and then obtain posterior distributions of the model parameters using MCMC techniques. Chapter 12 of Congdon (2005) illustrates how to fit Bayesian latent class models to data, particularly for the case when there is local dependence.

In classical inference the data are taken to be random, with the population parameters taken to be fixed, while in Bayesian inference the parameters themselves follow a probability distribution. This allows the consideration of models in the Bayesian paradigm that will otherwise will not be plausible in the classical framework because the Bayesian latent class model works by augmenting the data to produce a ‘known’ complete data likelihood. (Congdon, 2005, page 437) does go on to caution about the need to specify suitable priors to cope with model identifiability - i.e. to ensure that the resulting Bayesian model makes substantive sense.

3.13 Conclusion

The literature on capture-recapture is expansive and although it was historically confined to biological populations, the technique has currently been used in a wide range of applications from computer science to psychology to astronomy. In all these applications, the common strand is that underenumeration exists and needs to be estimated in order to gain an accurate representation of the population size. This chapter, has therefore sought to unify the existing methodology from a number of seminal sources and discuss their application to a triple system census, focusing on distinguishing between biases introduced into the population estimates through heterogeneity and dependence, and additionally quantify the level of over-enumeration.

In the 1991 UK census, the Census Validation Survey was used to estimate the level of underenumeration in the Census, but it was assumed that this survey was *perfect* at finding those individuals and households missed by the Census, which led to an under-estimation of the number of people missing (Brown et al. (1999)). In the 2001 census, a larger scale survey was implemented, and an extra assumption was made that individuals and households could be missed by this survey, known as the Census Coverage Survey, as well as by the Census. Further, if it is assumed that the probabilities of being counted by either the Census or Survey are *homogeneous* across the population, in a given post-stratum, and there is *independence* between the Census and Survey, then the estimate of those missed by both the Census and Survey can be found. By careful choice of factors known to affect an individual's probability of being counted, post-stratification can be used to split the population into sub-groups (post-strata), so that there is internal homogeneity within each sub-group. The independence assumption was difficult to guarantee, although there was operational independence between the Census and Survey processes (Brown, Abbott and Diamond (2006)). It is here that the thesis seeks to develop statistical methods that can adjust census counts for both underenumeration and overenumeration, in the presence of dependence through triple system estimation.

With information from three sources, it becomes possible to model the observed counts within a log-linear framework, and use the observed association patterns between the three lists to estimate the missingness. Overenumeration in the UK is generally perceived to be negligible in comparison to underenumeration; this was certainly a viable assumption under dual system estimation, but is clearly not true in triple system estimation. This is because the third, administrative list, is fraught with duplicates and so overenumeration does exist, and the objective is to quantify this. It has been shown in this chapter that it is possible to formulate a latent class log-linear model that can allow for the estimation of both the level of underenumeration and overenumeration. However, the issue of model identifiability imposes some restrictions on the model, that can be specified.

Chapter 4

Evaluation of Different Triple System Estimators

4.1 Introduction

The independence assumption in dual system estimation is heavily relied upon, but apart from Brown, Abbott and Diamond (2006) - and earlier by Bell (1993) in the US - there is currently no comprehensive work carried out to ascertain the performance of the dual system estimator in the presence of dependence. As it is not possible to estimate the level of dependence in a dual system framework directly, there is the need to use some ancillary information in order to test this. Accordingly, by means of information from administrative sources, Brown, Abbott and Diamond (2006) measured and adjusted for the dependence between the Census Coverage Survey and the Census counts in the UK 2001 census. They found that in most cases when coverage in both population counts are reasonably high the dual system estimation methodology was robust enough to cope with low levels of dependency.

However, an actual grasp of what constitutes 'low' and 'high' levels of dependency and 'low' and 'high' levels of coverage is something that has not been fully realised. Thus a simulation exercise was undertaken to seek to shed light on this, not just in the dual system (with data from the Census and Survey) but in a triple system framework when there is, in addition, data from a Third List as well as from the Census and Survey. The results of the simulation exercise are presented in this chapter. The chapter first outlines the motivation behind the Simulation Study and details how it was implemented and also looks at how the different estimators of the population fared in the presence of erroneous enumerations.

It must be noted that, in the chapter, it is assumed that the only dependence under consideration arises from the fact that the probability a person is counted or missed by a particular list is related to the probability that the same person is counted or missed on a different list. There is another source of dependence (not considered here) due to heterogeneity between individuals, who because of differences in their behavioural, demographic or other characteristics inherently, have different probabilities of being observed on any particular list. Also, it is assumed that the saturated model here is the one with no three-way interaction, i.e. the homogeneous association model. This is done for two reasons. Firstly, like the assumption of independence in a dual system setting, this assumption is required in order to be able to estimate the missing cell. Secondly, and more pertinently, by design the simulations were on the basis that the Third List was independent of the Census and Survey - implying that the observed cells contain sufficient information for the estimation of the population total. Thus it is possible to posit more complicated models (as will be shown), however these models will be anticipated to over-fit the observed data.

4.2 Simulation Study

The simulation study generated a population of 1000 individuals. Each individual was given a probability being counted in the Census, the Survey and the Third List. The individuals are then cross-classified into a 2x2x2 table according to their absence or presence on the three lists. Since the object is to find the individuals who fall into the $(0, 0, 0)$-cell corresponding to those missed in all three lists, attention is restricted to the incomplete table representing the individuals who are observed.

Different coverage probabilities are considered, and these take values of 30%, 50%, 70% and 90%. Now, since dual system estimation makes the assumption that there is no systematic relationship between the probability of an individual being counted in the Census and the same individual being counted in the Census Coverage Survey, the objective is to determine how robust a method it is to estimate the population size when some dependence is introduced. The dependence is represented by the odds ratio and took values in $\{1, 1.2, 1.4, 1.6, 1.8, 2\}$. Additionally, in order to investigate the performance of the population estimators at different odds ratios the reciprocals of the dependency were considered, i.e. in $\{\frac{1}{1.2}, \frac{1}{1.4}, \frac{1}{1.6}, \frac{1}{1.8}, \frac{1}{2}\}$.

Therefore, for given coverage and dependency levels a 2x2x2 contingency table is simulated on the basis of whether or not an individual is counted or missed in the Census, Survey and the Third List. Obviously since in reality the people missed by all three lists are unknown the n_{000} cell count is discarded, and the remaining seven cells are taken to be the 'observed' table of counts in the simulated population. The procedure is to then estimate the missing cell via the EM algorithm. The motivation for using the EM algorithm to find the missing cell count relies on the fact that the cell counts (the observed and unobserved) have some structure; so what is required is to find this structure. Using log-linear modelling, it becomes possible to posit different models depending on the perceived structure in the contingency table.

Since in a capture-recapture contingency table there is one cell missing by definition, the EM algorithm becomes useful. One often levelled criticism in capture-recapture modelling is that the estimation of the population size is based on a model that is deemed to closely fit the observed data from the incomplete contingency table which can be biased simply because the underpinning assumption is that the model describing the observed data also describes the unobserved individuals. Unfortunately there is no way this assumption can be checked, but it is fairly reasonable to assume, under the chosen circumstances, that this assumption holds.

Based on the best-fitting model to the observed cells in the contingency table, an estimate is found for the unobserved cell under this posited model. In essence the EM algorithm is used to estimate the unobserved cell such that the posited structure modelling the relationship between the observed cells remains the same. This is accomplished by the M-step using maximum likelihood estimation to fit the chosen log-linear model to the data, given an initial estimate of the number of people in the unobserved cell. Since none of the three lists achieve a 100% count of the population there is at least one person in the missing cell, but the EM algorithm starts the iterative process with an initial estimate of zero. The E-step then finds the conditional expectation under the model of the missing cell given the observed data, and this initial estimate. The M-step and the E-step are repeated giving new estimates of the unobserved cell until the change in the old and new estimates are infinitesimally small.

In the exercise three triple system estimators and the dual system estimator were considered. For each simulated data set, these four estimators were used to obtain the missing cell count, n_{000} and the total population size, N. The first triple system estimator considered is the mutual independence model (TSE1) which assumes that all three lists are independent of each other. It is important to see how this mutual independence model fares in comparison to the dual system estimator when there is some dependency. The second triple system estimator was the pairwise dependence model (TSE2). Here the model assumes that the Census and Survey are independent of the Third List. The third triple system model (TSE3) considered was the ‘saturated’ model, i.e. the homogeneous association model with all pairwise relationships between the Census, Survey and Third List present. Finally all three models were compared to the dual system estimator (DSE). Of interest was to determine if all the triple system estimators always outperformed the dual system estimator, regardless of the amount of dependence or the coverage probabilities. In order to assess the performance of each of these estimators, the bias and the standard error were calculated. The process was repeated multiple times[1] to yield the average bias and standard error.

On a cautionary note, the data has been simulated under dependence between the Census and Survey, and an assumption is made that bringing the Third List into the

[1] In most cases there were 10,000 iterations, but for a minority of cases this was computationally unfeasible and the number of iterations were reduced to 2,000.

frame does not introduce additional dependence. This assumption seems fairly reasonable in a UK context because the only feasible individual-level administrative list under consideration in a triple system scenario is the health records list. Further, the mechanism used to collate health data is sufficiently different to that used in the Census or Survey for it to be reasonably assumed that the Third List is independent of the Census or Survey. In other words, the coverage of an individual on the health register (referred to as the Third List) does not depend on the individual's coverage in the Census or Survey. This assumption will not strictly hold in other countries. For example, in the US, the administrative list used by Zaslavsky and Wolfgang (1990, 1993) was put together to better count those sub-populations who were difficult to enumerate in the Census and Survey. As such, in this context there is not only dependence between the Census and Survey, but the administrative list could be related to the Census, the Survey or both. However, the log-linear modelling framework proposed here is flexible enough to include additional dependence terms, if required (as shown in the next chapter).

As anticipated, the dual system estimator (DSE) is the most biased in all cases, when there is dependence and TSE2 and TSE3 are the least biased. However, TSE3 has larger standard errors and in some cases seems to over-estimate the population size. This is intuitive given that TSE3 is fitting the saturated model when a simpler model (with only the pairwise dependence between the Census and Survey) will suffice. It follows that any of the conditional independence models (i.e. the model with pairwise dependence terms between the Census and Survey and Census and List or the one with Census and Survey and Survey and List terms) may be unbiased, but will suffer from poorer precision when compared to the simpler model.

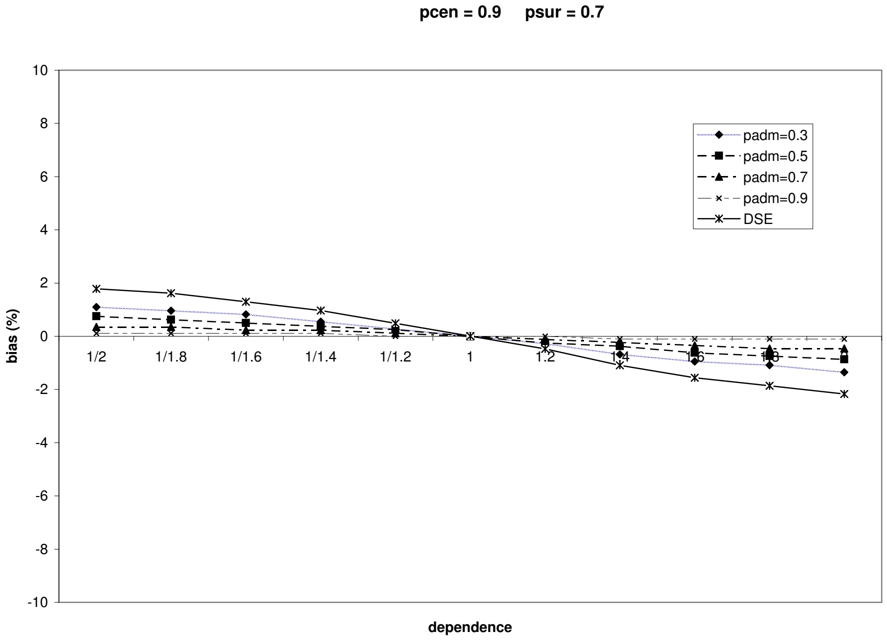
Given that TSE2 and TSE3 are virtually unbiased, by definition, the evaluation of the simulation exercise will be concentrating on TSE1 and how it performs for different levels of coverage and dependence. This is because it is imagined that the introduction of the Third List will improve the population estimates, but it is difficult to quantify how beneficial the Third List actually is. It becomes clear, however, that when there is high enough coverage on the Census and Survey, the Third List does not improve on the DSE a great deal, as shown in the plots below.

Figures 4.1 and 4.2 show how different coverage levels on the administrative list affect the simple triple system estimator that assumes independence between all three lists, TSE1. It is obvious that TSE1 is expected to be a biased estimator of the population size when there is some simulated dependence between the Census and Survey. This bias is positive when the dependence[2] $\gamma > 1$ and negative when $\gamma < 1$. In other words, when the bias is negative a person who is missed by the Census is more likely to be missed by the Survey. On the other hand, when the bias is positive then a person who is missed by the Census is more likely to be counted by the Survey. It is difficult to say which of the two is more likely to happen. Nevertheless, Figures 4.1 and 4.2 show that this bias is relatively

[2] where $\gamma = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}$

small (all the estimators have biases smaller in magnitude than 2.5%) when coverage is suitably high enough for the chosen range, $\{\frac{1}{2} < \gamma < 2\}$. On both graphs the bias under the DSE is plotted as well to give some idea as to the rewards of using the Third List. As intuitively expected, when there is independence all the estimators are unbiased, but assuming that the simulated dependence is 2 then the DSE will under-estimate the size of the missing population by a factor of 2. This follows considering that the estimate of the missing cell under DSE is given by $\hat{n}_{00} = \frac{n_{01}n_{10}}{n_{11}}$, when it should actually be $\hat{n}_{00} = \gamma\frac{n_{01}n_{10}}{n_{11}}$.

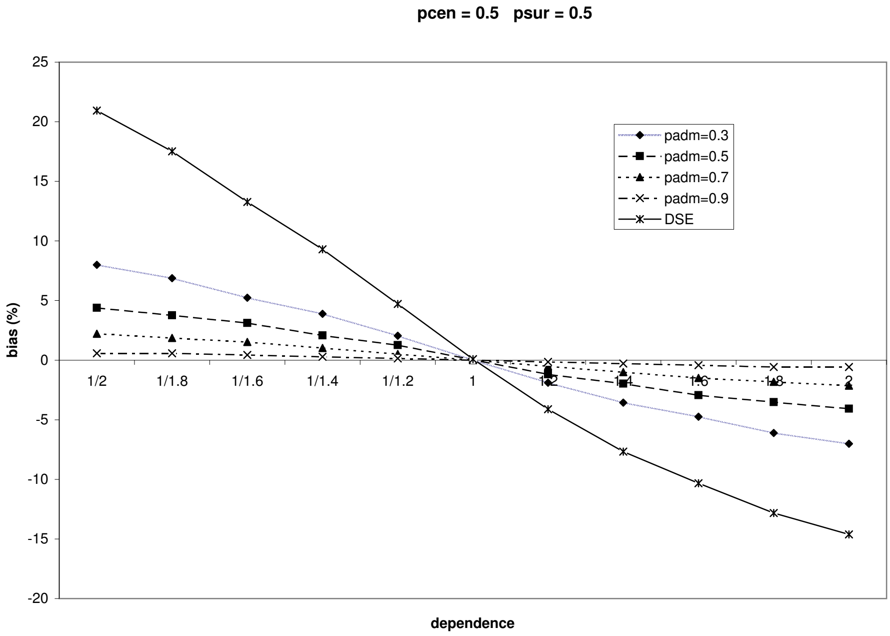
Figure 4.1: Performance of TSE1 for varying levels of Third List coverage.



Moreover, Figure 4.1 shows that the benefits of an administrative list when the Census achieves a population coverage of 90% while the Survey achieves 70% coverage (which is what was roughly achieved in the 2001 UK census) are relatively minimal in that the DSE in the most extreme case of dependence is actually relatively unbiased, with an absolute bias of 2.2%. However, the benefits of triple system estimation in the presence of dependence become clear in Figure 4.2 which show a somewhat significant reduction in the bias, even when the Third List only covers 30% of the population. Although it must be said that a poor covering administrative list becomes less useful when it is realised that, as is often the case in administrative records, there are erroneous enumerations. So the advantages of bringing in an administrative list that achieves poor population coverage are outweighed by the disadvantages due to the requirement to remove erroneous enumerations from the population estimate. The effect of erroneous enumerations on the population estimates are considered later on in the chapter.

For Figure 4.1 it is assumed that the coverage probability in the Census is 0.9 and the Survey coverage probability is 0.7. Here given $\gamma = 2$ the relative bias for the simulations

Figure 4.2: Performance of TSE1 for varying levels of Third List coverage.



when the Third List coverage probabilities are 0.3, 0.5, 0.7 and 0.9 are found to be −1.35%, −0.86%, −0.46% and −0.11%. By comparison the DSE has a relative bias of −2.17%.

For Figure 4.2 on the other hand, the Census and Survey coverage probabilities are both taken to be 0.5, and the relative bias when the simulated dependence is 2 for administrative list coverage levels of 0.3, 0.5, 0.7 and 0.9 are −7.02%, −4.07%, −2.14% and −0.57%. Apart from when the Third List has a 'poor' population coverage of 0.3, the absolute relative bias in all remaining cases is beneath 5%. This compares favourably to the relative bias for the DSE of −14.62%.

The presence of the administrative list does improve on the population estimate; more so, this improvement can be shown to be particularly significant when the Census and Survey fail to achieve reasonable population coverage. When there is 50% coverage in the Census and Survey, the DSE bias is −14.62% for $\gamma = 2$ and 20.93% for $\gamma = \frac{1}{2}$. However, the bias for an administrative list with coverage of 30% is −7.02% and 8.00%, respectively, which is roughly equivalent to a two-thirds reduction in bias when $\gamma = 2$ and a half for $\gamma = \frac{1}{2}$. Furthermore, increasing the administrative list coverage to 50%, leads to a bias of −4.07% and 4.39% (for $\gamma = 2$ and $\frac{1}{2}$), which is almost a 50% improvement on the TSE bias and roughly 80% on the DSE bias.

Another observation from the simulation results concerns the symmetry. One of the reasons behind choosing reciprocals was to look at the behaviour about dependence of 1 (i.e. independence) since there is no bias when $\gamma = 1$ but the bias is positive for γ between (0, 1) and negative for γ between (1, ∞). The relative bias for $\gamma = \frac{1}{2}$ in Figure 4.1 when the Third List coverage was 0.3, 0.5, 0.7 and 0.9 was respectively 1.10%, 0.75%, 0.35%

and 0.11%. The DSE bias was 1.79%. However, the relative bias for dependence, γ, of 2 is -1.35%, -0.86%, -0.46% and -0.11%, with a DSE bias of -2.17%. This shows that there is symmetry for high coverage probabilities, but this symmetry diminishes as the coverage probability drops. This asymmetry is more evident in the dual system estimator, as shown in Figure 4.2 where the relative biases under TSE1 at 2 and $\frac{1}{2}$ are -0.59% and 0.59% when the administrative List coverage is 0.9; -2.14% and 2.21% when the List coverage is 0.7; -4.04% and 4.39% when the List coverage is 0.5; and -7.02% and 8.00% when the List coverage is 0.3. For the DSE the relative bias is 20.93% when the dependence is $\frac{1}{2}$ compared to -14.62% when the dependence is 2. The lack of symmetry is more apparent when looking at higher levels of dependence. Table 4.1 shows the results of the relative biases when there is a simulated dependence between the Census and Survey of $\frac{1}{8}$ and 8, and some asymptotic properties of the bias of the DSE are presented below. Trivially, it may be observed that the TSE biases are bounded by the DSE.

Table 4.1: Relative bias at simulated dependence of $\frac{1}{8}$ and 8

	Dependence	
	$\gamma = \frac{1}{8}$	$\gamma = 8$
pcen=0.5, psur=0.5		
padm=0.3	25.440%	-17.387%
padm=0.5	13.217%	-10.997%
padm=0.7	6.287%	-5.760%
padm=0.9	1.735%	-1.691%
DSE	91.718%	-32.233%
pcen=0.9, psur=0.7		
padm=0.3	2.366%	-3.981%
padm=0.5	1.513%	-2.580%
padm=0.7	0.811%	-1.484%
padm=0.9	0.214%	-0.409%
DSE	3.976%	-6.389%

The limiting behaviour of the dual system estimator can also be investigated to give some indication of how the triple system estimators behave since Figures 4.1 and 4.2 show that the triple system estimator biases lie within the dual system estimator bias. This is reasonable in view of the fact that the dual system estimator is broadly not as efficient as the triple system estimators. Given that the triple system estimators are complicated functions of γ, it is not easy to ascertain how the different TSEs change with varying dependencies and coverage probabilities. However, simple expressions can be found for the DSE, at varying levels of dependence and coverage. Furthermore, since it has been shown that the DSE bounds the TSEs, obtaining expressions of how the DSE behaves as γ tends to zero and infinity, does provide some information as to the asymptotic behaviour of the TSEs.

Recall that in Chapter 3, the Census and Survey coverage probabilities were defined to be π_{1+} and π_{+1}, and the dependence, γ, between the two lists is $\gamma = \frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}}$. It was shown that another expression for γ is

$$\gamma = \frac{\pi_{11}\left(1 - \left(\pi_{1+} + \pi_{+1} - \pi_{11}\right)\right)}{\left(\pi_{1+} - \pi_{11}\right)\left(\pi_{+1} - \pi_{11}\right)}. \tag{4.1}$$

Now, (4.1) can be expressed in terms of a quadratic function of π_{11},

$$\pi_{11}{}^2\left(1-\gamma\right) + \pi_{11}\left(1 - \pi_{+1} - \pi_{1+} + \gamma\left(\pi_{+1} + \pi_{1+}\right)\right) - \gamma\pi_{1+}\pi_{+1} = 0.$$

Therefore, supposing the coverage probabilities π_{1+} and π_{+1} and dependence, γ, are known, then

$$\pi_{11} = \frac{-\left(1 - \pi_{+1} - \pi_{1+} + \gamma\left(\pi_{+1} + \pi_{1+}\right)\right) \pm \sqrt{\left(1 - \pi_{+1} - \pi_{1+} + \gamma\left(\pi_{+1} + \pi_{1+}\right)\right)^2 + 4\left(1-\gamma\right)\left(\gamma\pi_{1+}\pi_{1+}\right)}}{2\left(1-\gamma\right)}. \tag{4.2}$$

After obtaining the value of π_{11}, the rest of the probabilities in the contingency table, Table 4.2, can be found since the marginal probabilities, π_{1+} and π_{+1} are already known. In view of the fact that π_{11} can be expressed as a function of the dependence, γ, it becomes possible to ascertain the limiting behaviour of the relative bias as γ tends to zero and infinity. It must also be noted that even though there is some dependence between the Census and Survey, these two are assumed to be independent of the Third List.

Table 4.2: Probabilities from 2x2 contingency table

π_{11}	π_{01}	π_{1+}
π_{01}	π_{00}	(1-π_{1+})
π_{+1}	(1-π_{+1})	

Now, for the case when the Census (π_{1+}) and Survey (π_{+1}) coverage are respectively 90% and 70%, then Equation (4.2) simplifies to

$$\pi_{11} = \frac{0.6 - 1.6\gamma \pm \sqrt{\left(1.6\gamma - 0.6\right)^2 + 4\left(1.6\gamma\right)\left(1-\gamma\right)}}{2\left(1-\gamma\right)}. \tag{4.3}$$

Similarly when the Census and Survey coverage probabilities are both 50%, then

$$\pi_{11} = \frac{-\gamma \pm \sqrt{\gamma^2 + 4\gamma\left(1-\gamma\right)}}{2\left(1-\gamma\right)}. \tag{4.4}$$

What happens to π_{11} as γ tends to zero?

As the dependence between the Census and Survey becomes smaller,

$$\begin{aligned} \lim_{\gamma\to 0}\pi_{11} &= 0.6 \qquad \text{for } \pi_{1+} = 0.9 \text{ and } \pi_{+1} = 0.7 \\ &\text{or} \\ &= 0 \qquad \text{for } \pi_{1+} = 0.5 \text{ and } \pi_{+1} = 0.5. \end{aligned}$$

Generally, for any set of marginal probabilities $\{\pi_{1+}, \pi_{+1}\}$ as γ tends to zero, the algebraic limit of (4.2) simplifies to

$$\begin{aligned}\lim_{\gamma \to 0} \pi_{11} &= \frac{(1 - \pi_{1+} - \pi_{+1}) \pm (1 - \pi_{1+} - \pi_{+1})}{-2} \\ &= \max\{0, (\pi_{1+} + \pi_{+1} - 1)\}\end{aligned}$$

since $\pi_{11} \geq \max\{0, (\pi_{1+} + \pi_{+1} - 1)\}$.

Now having found the probability of being counted by both, π_{11}, the other three cell probabilities can be found using the marginal probabilities, π_{1+} and π_{+1}. Thus for $\pi_{1+} = 0.9$ and $\pi_{+1} = 0.7$ then $\pi_{11} = 0.7$, the four cell probabilities are $\pi_{11} = 0.6$, $\pi_{10} = 0.3$, $\pi_{01} = 0.1$ and $\pi_{00} = 0$. Also, for $\pi_{1+} = 0.5 = \pi_{+1}$, $\{\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\} = \{0.0, 0.5, 0.5, 0.0\}$. Using the fact that the Third List is independent of both the Census and Survey then $\pi_{ijk} = \pi_{ij+}\pi_{++k}$. So assuming the Third List has coverage of 70%, the resulting eight cell probabilities for the case when the Census and Survey coverage is 90% and 70% are $\{0.420, 0.180, 0.210, 0.090, 0.070, 0.030, 0.000, 0.000\}$. For a population of 1000 people the estimate of the population under the DSE is given by $\frac{n_{1++}n_{+1+}}{n_{11+}} = \frac{(420+180+210+90)\times(420+180+70+30)}{(420+180)} = 1050$. The relative bias is therefore 5%. Conversely for Census and Survey coverage of 50% and Third List coverage of 70%, the ensuing eight cell counts are $\{0, 0, 350, 150, 350, 150, 0, 0\}$. Since $n_{11+} = 0$, the 'usual' DSE does not work as it gives an undefined estimate, and so the Chapman corrected dual system estimator (Chapman (1951))

$$\hat{N}^C = \frac{(n_{1++} + 1)(n_{+1+} + 1)}{(n_{11+} + 1)} - 1 \tag{4.5}$$

is used. This yields a population estimate of 251,001 and a relative bias of 2,500%.

What happens to π_{11} as γ tends to infinity?

As the dependence increases, then it can be demonstrated that

$$\begin{aligned}\lim_{\gamma \to \infty} \pi_{11} &= 0.7 \qquad \text{for } \pi_{1+} = 0.9 \text{ and } \pi_{+1} = 0.7 \\ &\text{or} \\ &= 0.5 \qquad \text{for } \pi_{1+} = 0.5 \text{ and } \pi_{+1} = 0.5.\end{aligned}$$

In general for any given marginal probabilities π_{1+} and π_{+1}, as γ tends to infinity the expression (4.2) simplifies to

$$\begin{aligned}\lim_{\gamma \to \infty} \pi_{11} &= \frac{(\pi_{1+} + \pi_{+1}) \pm (\pi_{1+} - \pi_{+1})}{2} \\ &= \min\{\pi_{1+}, \pi_{+1}\}\end{aligned}$$

since $\pi_{11} \leq \min\{\pi_{1+}, \pi_{+1}\}$.

So for the case with Census coverage of 90% and Survey coverage of 70%, then $\pi_{11} = 0.7$, $\pi_{10} = 0.2$, $\pi_{01} = 0.0$ and $\pi_{00} = 0.1$. Further using the independence of the Third List, and assuming coverage of 70%, the eight cell probabilities are $\{0.49, 0.21, 0.14, 0.06, 0.00, 0.00, 0.07, 0.03\}$. Under the DSE with 1000 people the estimated population is 900, leading to a bias of -10%. Likewise, for Census, Survey and Third List coverage of respectively 50%, 50% and 70%, then $\pi_{11} = 0.5 = \pi_{00}$ and $\pi_{10} = 0 = \pi_{01}$. Accordingly the eight cell counts assuming a population of 1000 people is $\{350, 150, 0, 0, 0, 0, 350, 150\}$. This gives a DSE of 500, and a relative bias of -50%.

Lastly, it is evident from the limiting behaviour of the DSE relative bias that there is a lack of symmetry as the dependence, γ, gets larger or smaller. This has been demonstrably shown by the fact that the lower and upper limits of the relative bias of the DSE are (-5%, 10%) for 90% Census coverage and 70% Survey coverage, and (-50%, 2500%) for Census and Survey coverage of 50%.

The preceding discussion has investigated the bias of the different population estimators. However, the bias is concerned with how accurate the estimator is in measuring the quantity of interest, and is just one measure of an estimator's performance. The variance is another measure which looks at how precise this estimator is. Clearly, an estimator may be precise but inaccurate and vice versa. Thus to determine which of the estimators was the best the mean squared error can be used (Cox and Hinkley, 1974, page 253). In essence, the mean squared error rewards small biases but penalises larger standard errors. It is therefore a useful tool in comparing the performance of the dual and triple system estimators. In an ideal world, the best estimator will have the lowest bias and the lowest variance. The simulation exercise showed that on the one hand though TSE3 is unbiased it comes with large standard errors, while on the other hand TSE1 has some bias, but the standard errors may be small for some cases. So the objective is to compare which of these estimators performs the best, under different scenarios.

Figures 4.3 and 4.4 compare the mean squared error for the different triple system estimators, under the scenarios detailed above when the Census and Survey coverage probabilities of 0.9 and 0.7, and 0.5 and 0.3. The first thing of note is that the mean squared error for TSE3 is larger than the respective mean squared errors for TSE1 and TSE2, which supports the assertion that TSE3 is an inefficient estimator. Although, TSE3 like TSE2 is relatively unbiased, the associated large variance of the estimator has the effect of inducing a high mean squared error, in comparison with the biased but low variability TSE1. When the coverage in the Census, Survey and Third List are high then there is very little to distinguish between the three estimators, in terms of their mean squared error. It can also be noticed that as the Third List coverage probability increases there seems to be very little difference between the mean squared error plots for TSE1 and TSE2. Additionally, as the Census and Survey coverage get higher, it appears that TSE3 becomes pejoratively less efficient when compared to TSE1 and TSE2.

In both figures the dual system performance is included, and from Figure 4.3 it can be seen that when the Census and the Survey respectively cover 90% and 70% of the population but the Administrative List is poor (at 30%), then the DSE copes well with

Figure 4.3: Comparison of the root mean square error for different triple system estimators and the dual system estimator (for p_{cen}=0.9 and p_{sur}=0.7).

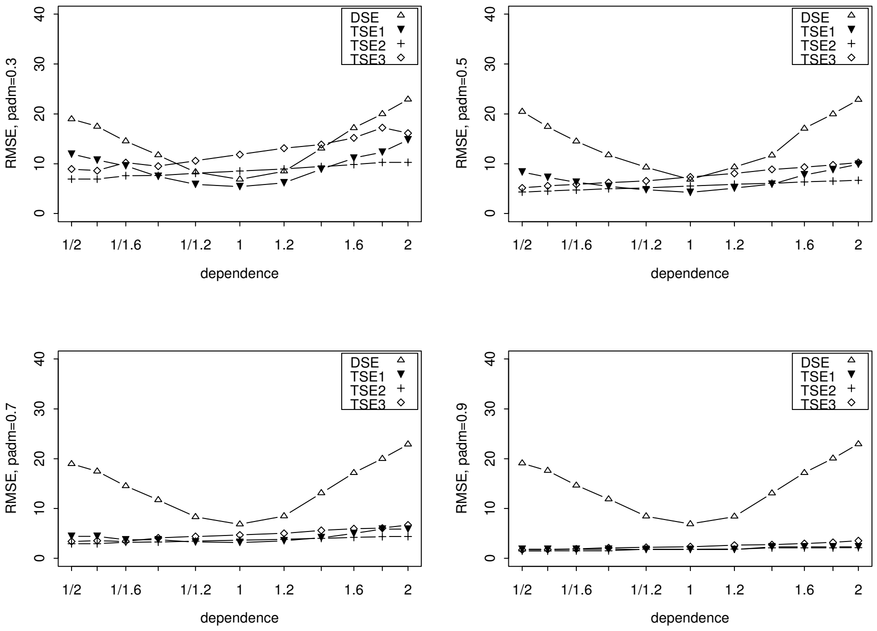
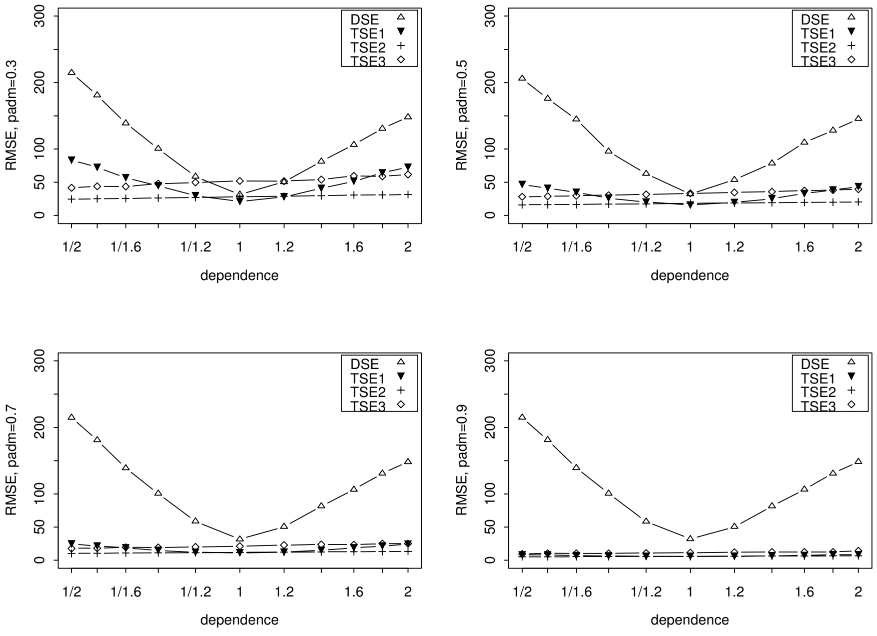


Figure 4.4: Comparison of the root mean square error for different triple system estimators and the dual system estimator (for p_{cen}=0.5 and p_{sur}=0.5).



failures of the independence assumption. Here the DSE, although slightly biased has the lower RMSE than the two unbiased TSEs. Indeed, for $\frac{1}{1.2} < \gamma < 1.2$ the DSE is a better estimator, and for $\frac{1}{1.4} < \gamma < 1.4$ the biased triple system estimator, TSE1, is the most efficient. Even when the Census and the Survey do not achieve a decent coverage of the population (i.e. both have 50% coverage) and the Third List achieves 30%, TSE1 is the most efficient estimator of the population for $\frac{1}{1.2} < \gamma < 1.2$. From both figures, it can be inferred that only at 50% coverage of the Administrative List is the DSE less efficient than all the TSEs at all levels of dependence. Indeed, the bottom right plots in both figures representing an Administrative List coverage of 90% show that the mean squared error of all the TSEs is much lower than that of the DSE.

There is some degree of reasonableness to these results. For the case when both the Census and Survey achieve high coverage of the population, the Administrative List does not have that many people to find (for a population of 1000 people, with 90% Census coverage and 70% survey coverage, even with dependence of 2, there are only 44 people expected to be found, whereas with 50% coverage there are now 293 people[3].

What the mean squared error plots are crudely saying is that the simplest triple system estimator (i.e. TSE1) performs reasonably well as it has the lowest root mean squared error for all dependency levels - despite the fact that TSE1 is biased, the variance is not comparatively smaller than the other less biased estimators. Further, even though the best model is TSE2, TSE1 is consistent enough as an estimator of the population size to merit consideration. There could be an argument to always fit TSE3 to the data since it has an easy close-form expression, i.e. $\hat{n}_{000} = \frac{n_{111}n_{100}n_{010}n_{001}}{n_{110}n_{101}n_{011}}$. However, the above simulation exercise has shown that even when there is relatively high dependence between the Census and Survey, doing this is not the most efficient way of determining the missing cell. A better way is to fit TSE1, i.e. $\hat{n}_{000} = \frac{\hat{n}_{0++}\hat{n}_{+0+}\hat{n}_{++0}}{\hat{N}^2}$ which unfortunately does not have a closed-form solution.

Nonetheless, this result becomes useful because TSE1 can be re-written in an alternative way when it is noticed that mutual independence is synonymous with all pairwise independence, so the probability of being found in the $(i,j,k)^{th}$ cell in the 2x2x2 contingency table is given by

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}.$$

This can be re-written as

$$\frac{\hat{n}_{ijk}}{\hat{N}} = \frac{\hat{n}_{i++}\hat{n}_{+j+}\hat{n}_{++k}}{\hat{N}^3},$$

and so the estimator for the population size under mutual independence becomes

$$\hat{N} = \sqrt{\frac{\hat{n}_{i++}\hat{n}_{+j+}\hat{n}_{++k}}{\hat{n}_{ijk}}}. \tag{4.6}$$

Now, since the marginal sums are sufficient statistics under mutual independence, the unknown population size can be thought of in terms of only the individuals who were counted at least once on any of the lists,

$$\bar{\hat{N}} = \sqrt{\frac{n_{1++}n_{+1+}n_{++1}}{n_{111}}}. \tag{4.7}$$

[3] for $\gamma = 2$, $\hat{n}_{00+} = 44$ when $p_{cen} = 0.9$, $p_{sur} = 0.7$ and $\hat{n}_{00+} = 293$ for $p_{cen} = p_{sur} = 0.5$

Equation (4.7) is an interesting triple system estimator, as it implies that when it is believed that there is mutual independence between the three counts of the population then the unknown population size can be estimable by only using the observed marginal totals, n_{1++}, n_{+1+} and n_{++1}, which is admittedly much more simpler than the previous mutual independence estimator (for which it must be remembered no closed-form solution exists). The obvious advantage of this estimator unlike the maximum likelihood estimator $\hat{N}$, is that $\bar{\hat{N}}$ has a closed solution and so can be evaluated without the need for a log-linear model, or the use of the EM algorithm. Darroch (1958) mentions that the above estimator (which will be referred to as the naive TSE1) is not the maximum likelihood estimate of the population size. However, is it an 'unreasonable' estimator?

An investigation was, therefore, undertaken to ascertain how differently the 'naïve' mutual independence TSE fared when compared to the mutual independence TSE that was found by maximizing the likelihood. The results of which are presented in Figures 4.5 and 4.6. Again, when coverage in all the three lists is high enough (above 70%) then the 'naïve' estimator is not that biased when compared to the ML estimator. Also, as the dependency term, γ moves away from 1 (i.e. independence) the bias in the 'naïve' estimator becomes greater. Figure 4.6 shows, however, that when the coverage is poor and there is dependence then the triple system estimator that fails to account for this dependence will lead to a biased estimate of the population size. In fact, at high levels of dependence and low coverage levels this naïve TSE is almost as bad as the DSE. Another observation is that the 'naïve' estimator can give negative values of the missing cell, unlike the MLE. As a demonstration, supposing the census and survey coverage is 50%, but the administrative coverage is 90%, and $\gamma = 2$, then the seven observed cells are $\{21, 21, 29, 264, 186, 186, 264\}$, and there are 971 people observed in total. For a simulated population of 1000, the correct estimate of the missing n_{000} is 29 people. Yet, the estimate of the total population under the 'naïve' estimator, $\bar{\hat{N}}$, is 923, which gives an estimate of $n_{000} = -48$. In contrast, under the MLE, $\hat{N} = 994$, which although biased gives a positive estimate of n_{000}. This is a flaw in the 'naïve' estimator similar to how the DSE behaves when there is poor coverage in the Census and Survey but good coverage on the Administrative List.

Figures 4.7 and 4.8 plot the standard errors of the 'naïve' and maximum likelihood independence triple system estimators. For both plots the standard errors of the maximum likelihood TSE are always lower than those of the 'naïve' TSE, with the ML TSE standard errors being roughly between a half to a quarter of the 'naïve' TSE. This highlights the fact that although the biases of the 'naïve' and ML independence triple system estimators may be roughly similar, the spread of the 'naïve' TSE is much wider than the MLE.

So the point to make here is that though $\{n_{i++}, n_{+j+}, n_{++k}\}$ may be sufficient statistics of the mutual independence model, $\{n_{1++}, n_{+1+}, n_{++1}\}$ are unfortunately not. Indeed since the variance of the MLE is much smaller than the 'naïve' estimator it can be concluded that this 'naïve' TSE, although simple to calculate is not very efficient.

Figure 4.5: Relative Bias of the Mutual Independence Triple System Estimators (for p_{cen}=0.9 and p_{sur}=0.7).

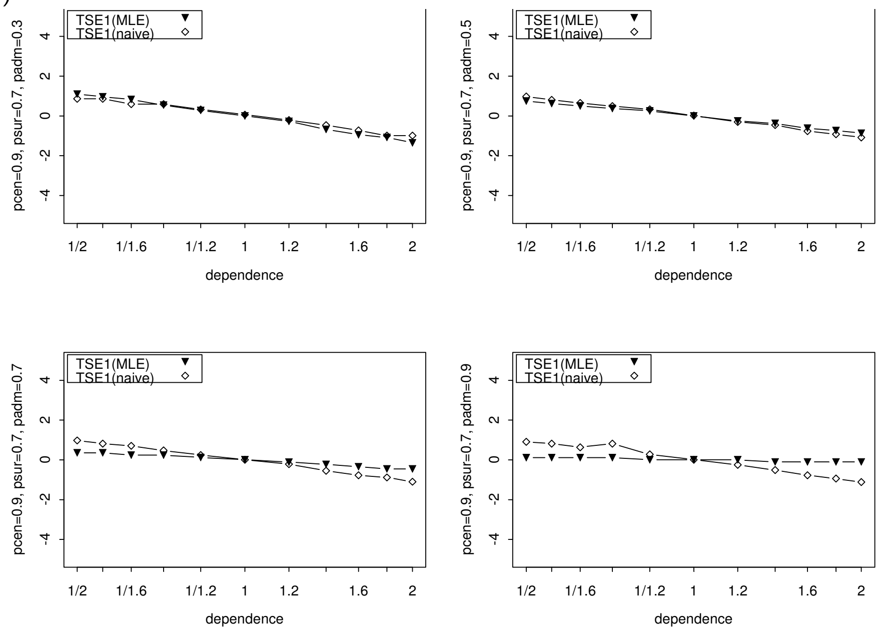
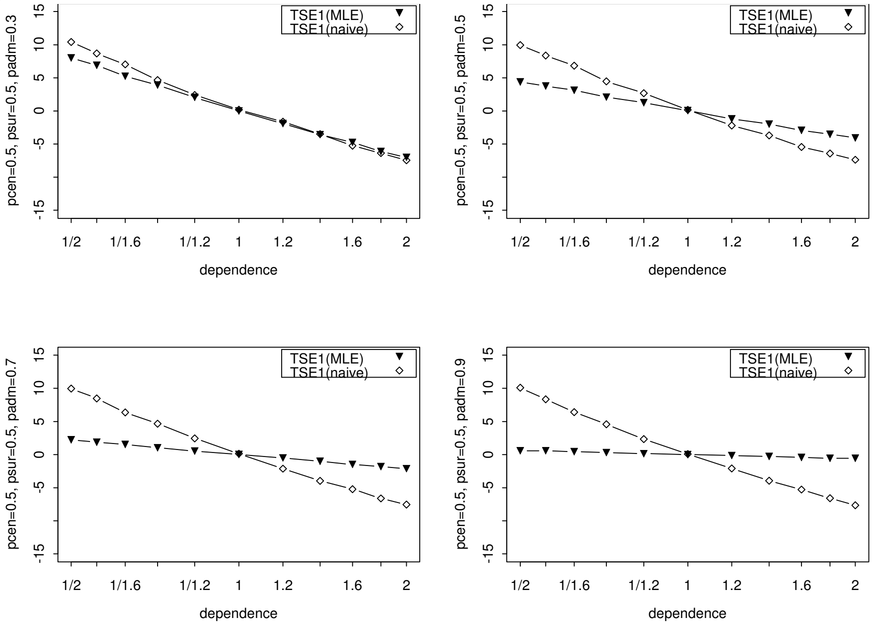


Figure 4.6: Relative Bias of the Mutual Independence Triple System Estimators (for p_{cen}=0.5 and p_{sur}=0.5).



Nevertheless, in practical terms, the 'naïve' estimator does merit consideration, particularly when there is high coverage in all three lists, as it has been shown to be almost unbiased. Finally, there are a number of ways in which the variance can be found. For the simulation exercise, it was possible to obtain the variance either empirically (directly from the results of the simulations) or by employing asymptotic formulae (given in Chapter 3). These asymptotic results rely on there being a large enough sample for asymptotic theory to hold. However, there has been some concern raised by some authors that in capture-

Figure 4.7: Standard Errors of the Mutual Independence Triple System Estimators (for p_{cen}=0.9 and p_{sur}=0.7).

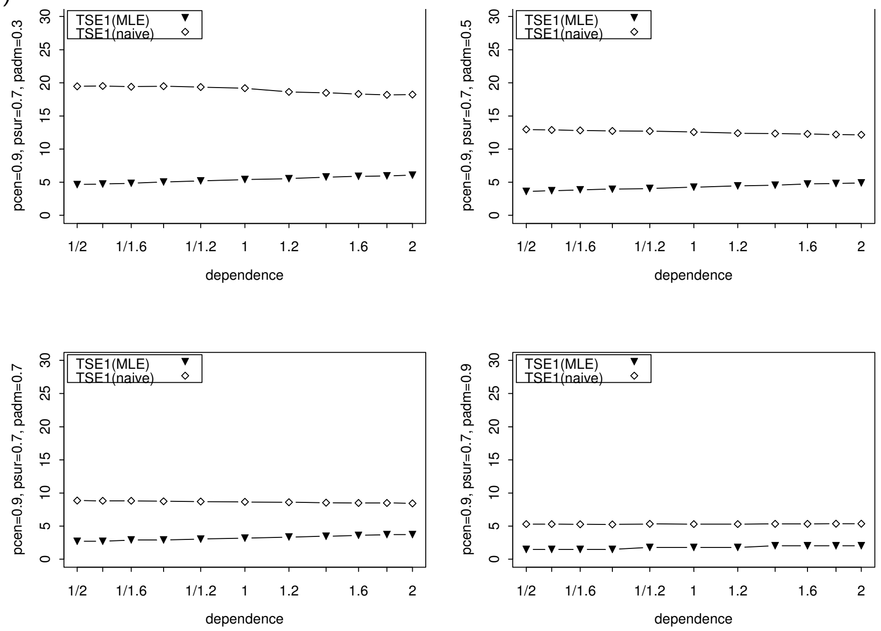
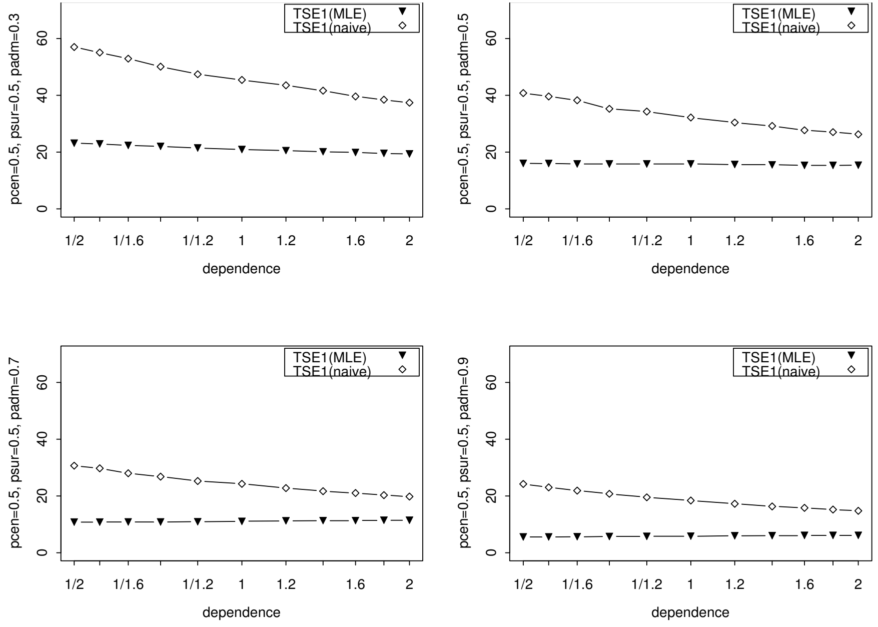


Figure 4.8: Standard Errors of the Mutual Independence Triple System Estimators (for p_{cen}=0.5 and p_{sur}=0.5).



recapture, the assumption of normality could be unrealistic, since in reality there is a relatively flat likelihood surface leading to some positive skewness (Coull and Agresti (1999), (Agresti, 2002, page 513)).

Figures 4.9 and 4.10 show that the asymptotic and empirical standard errors are similar for all three triple system estimators and the dual system estimator. This goes to demonstrate that, in the simulations, the assumption of normality is reasonable. So albeit the normal approximation may have some limitations, it was found that for the Simulation Study, this approximation did not have a negative influence on the calculation of the variance. However, it can be noticed that for the dual system estimator the differences between the asymptotic and empirical standard errors become more pronounced the further away from independence. This is another intuitive result stemming from the assumptions under which the asymptotic variance of the dual system estimator is calculated by using the Delta method. The Delta method for the dual system estimator asymptotic standard error relies on the cell counts being independent binomially distributed, which is perfectly valid under independence but not necessarily so when there is dependence between the Census and Survey. Nonetheless, the estimated precision of the dual system estimator using the asymptotic methods is reasonably good, and again this is particularly true when the coverage levels are high. Also, the DSE standard errors get larger as $\gamma \to 0$, whereas the TSE standard errors get larger as $\gamma \to \infty$.

Furthermore, Figures 4.9 and 4.10 show that TSE1 has lower standard errors than the other estimators. So it can be concluded that despite TSE1 being relatively more biased than the other two triple system estimators considered during the simulation, the variance of TSE1 seems sufficiently smaller than the TSE2 and TSE3. Hence, the reason why the MSE plots (see Figures 4.3 and 4.4) illustrate that TSE1 has a lowest mean squared error in most cases. Indeed, when the Third List has poor coverage (i.e. 30%) the standard errors of the (biased) DSE are much lower the (unbiased) TSE3, which fits the saturated model to the data. This implies that on average, the collection of estimates of the population total under the DSE are closer to the true population total than the estimates found using TSE3, even when there is some dependence between the Census and Survey. In conclusion, it appears that at moderate levels of dependence, the independence model (TSE1) is the most efficient. This is because, the dependence model (TSE2) and the ‘saturated’ homogeneous association model (TSE3) although unbiased are susceptible to high variability.

As a last point, there is a school of thought that suggests a move away from the census (here, referring to the traditional method of enumeration), with more of an emphasis on administrative records owing to the challenges of achieving accurate coverage of the population (Keohane (2008)). If the Administrative List has a comparatively higher coverage of the population than the Census, then it might be better for it to replace the Census. There is also a definite case to be made for the argument that an Administrative List to be more likely to be independent of the Survey or Census. So dual system estimation could be employed, with the Administrative List replacing the traditional census enumeration as the primary source. The major shortcoming of this argument lies in the existence of erroneous counts found on the Administrative List, which currently cannot be easily removed. Thus, unless there is an explicit adjustment for the overenumeration that results from using an Administrative List it will be difficult to produce geographically-accurate population estimates.

Figure 4.9: Comparison of the standard errors of the dual and triple system estimators (for p_{cen}=0.9 and p_{sur}=0.7).

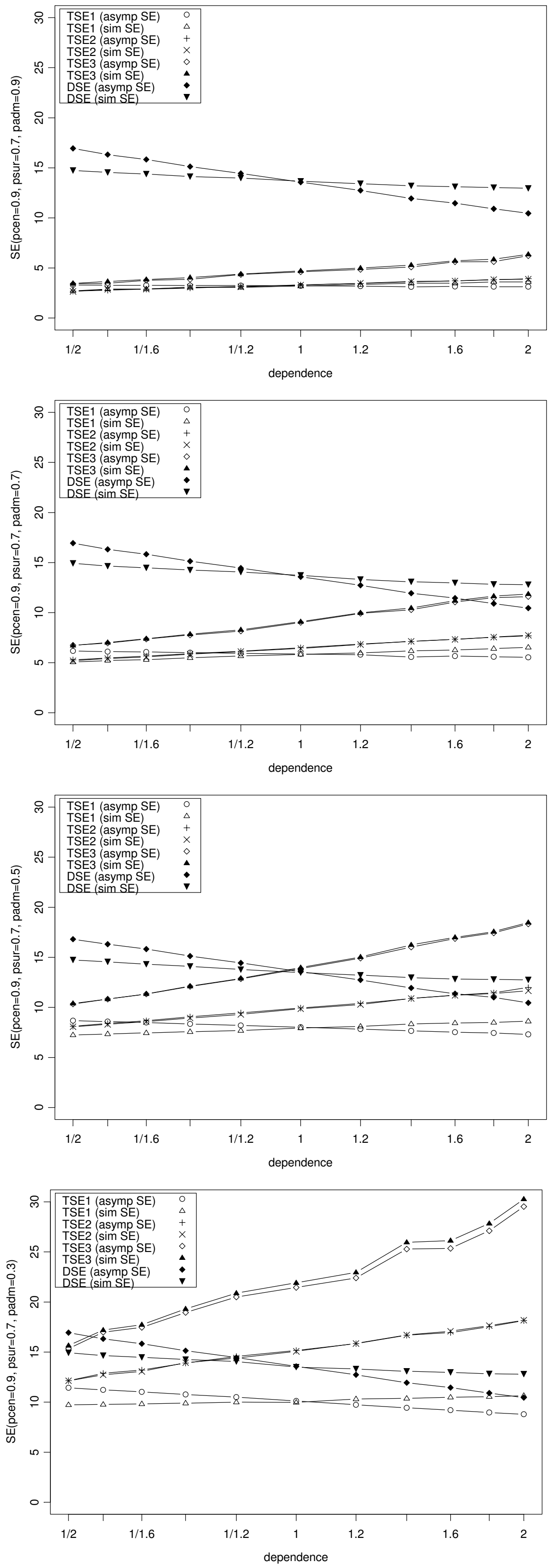
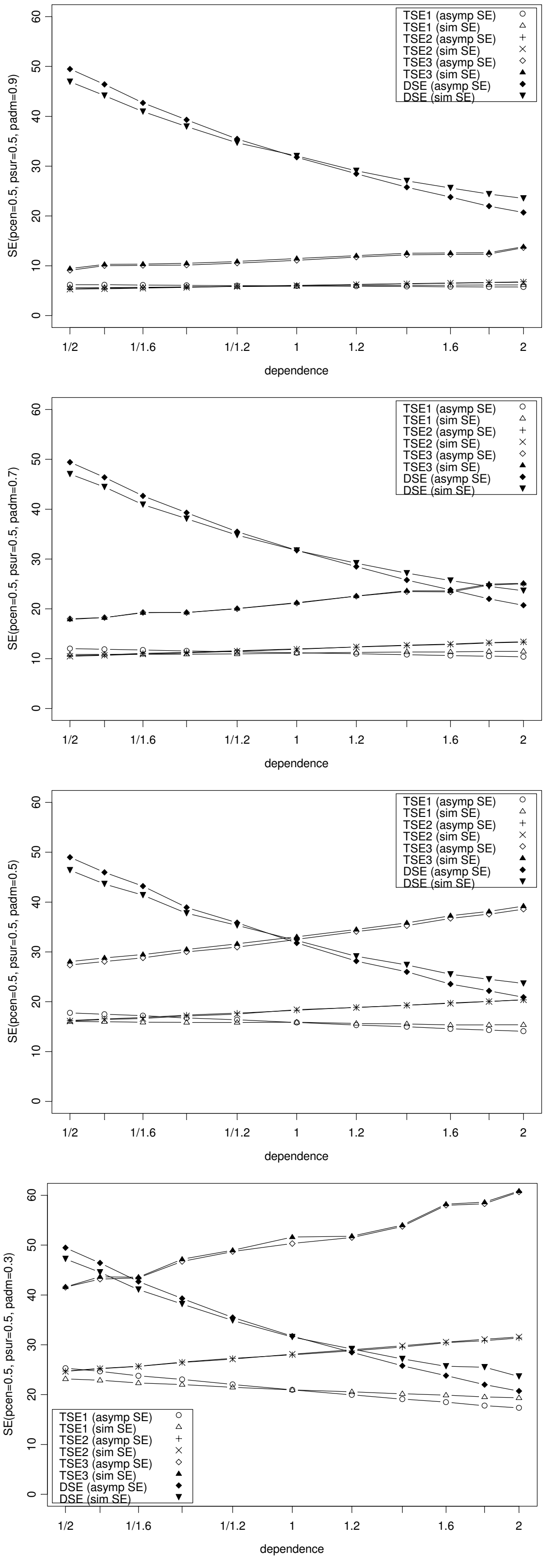


Figure 4.10: Comparison of the standard errors of the dual and triple system estimators (for p_{cen}=0.5 and p_{sur}=0.5).



4.3 Impact of Erroneous Enumerations on the Estimators

4.3.1 Erroneous Enumerations only in the Third List

The simulation results presented thus far were based on the assumption that there were no erroneous counts on any of the three lists. Whilst it may be reasonable to assume that the Census and Survey clerical matching and checking procedures sufficiently removed any fictitious enumerations and duplicates, it is difficult to allow this assumption for the administrative list. One overarching factor lies in the definition of the 'usual resident' population, which although will be the same for the Census and the Survey, may be different for the Third List. As such, though the population counts found by the Census and Third List at an aggregate level will be similar, when looking at a sub-aggregate level there will be differences which may have an impact. Health records, which are currently the most feasible administrative list that can be used as the Third List in the UK, suffer from overenumeration due to their failure to remove people who have moved. In 2001, the Survey counted people at their 'usual residence'. Moreover, the enumerators collected additional information on other possible locations where individuals may have been enumerated in the Census. If the same takes place in 2011 for the Census and Survey, but the Third List is brought in so as to implement triple system estimation, then there needs to be an explicit adjustment for overenumeration.

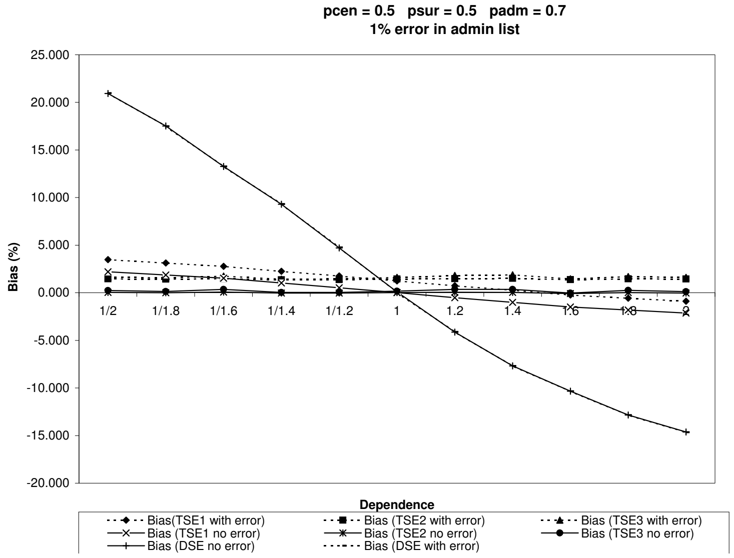
The 2001 census made determined efforts to clearly enumerate movers in the correct place. This was because the treatment of movers becomes important especially when considering how operational independence between the Census and Survey was implemented in the One Number Census design. If the first and second counts of the population are to be carried out in accordance with capture-recapture methodology in the strictest sense, then the Census and Survey will have to take place on the same day. However for operational independence to be achieved, it is not possible for the Census and Survey to both be in field at the same time. Thus in 2001 the Survey was carried out roughly three weeks after the Census was completed. It becomes inevitable that people will move in or out of the sampled areas between the time the Census was carried out and the Survey interviewers went into the field. Consequently, there was a procedure where some information was collected on out-movers and in-movers. This information was then matched to enable each individual to be counted at one, and only one, address. Although this proved to be fairly successful, when there are three lists this becomes extraordinarily difficult to implement. More so, given that there is a lag between the time people move and when the Third List is updated to take account of this move, the problem of people enumerated in the wrong address is exacerbated in triple system estimation.

Bearing this in mind, the next part of the simulation exercise therefore looked at the effect of erroneous enumerations on the different estimators. For dual system estimation, it is envisaged that the data matching and processing of the Census and Survey will remove any duplicates or fictitious people, and so every individual in the resulting 2x2 table can be considered to be 'real'. This is because, for the sample of areas in the Survey, any incorrect, fictitious or otherwise erroneous data are duly removed before producing the table of people missed or counted in the Census or Survey. When information from the Third List is used, there will be both 'erroneous' and 'real' individuals in the resulting 2x2x2 table. Mostly the erroneous individuals will be people who are counted in the wrong location, for example students, children of divorced parents, or highly mobile young people. So in the design of the simulation exercise another assumption was initially made. Given that the 2x2 table representing people counted or missed by the Census and Survey has been cleaned of any erroneous enumerations, the only potential source of erroneous enumerations should be realistically speaking through the Third List. Subsequently, an assumption was made that erroneous enumerations will be found only in the cell count representing those people who are counted on the Third List but missed by both the Census and Survey, i.e. in the (0,0,1)-cell. For that reason, erroneous counts were added to the (0,0,1)-cell, and the different population estimators - DSE, TSE1, TSE2 and TSE3 - were re-calculated.

For each simulation, the erroneous enumerations were added to the (0,0,1)-cell simulated under a N(10, 2) distribution, rounded to the nearest integer. The choice of this distribution was arbitrary but it was motivated by considering the currently available health registers. Since the population constituted of 1000 people, it was decided that, prior to carrying out the triple system estimation, the matching process using statistical and computer matching software has been able to remove most of the erroneous enumerations. So what remains are the people who have been enumerated in the wrong location, which should, realistically speaking, not be a substantial proportion of the population. As a result, for the population being considered it was decided that roughly 10 people are erroneous (representing 1% of the population). As a proportion of the simulated population total these erroneous enumerations maybe considered negligible; they do however make a fairly significant proportion of the (0,0,1)-cell count.

This is particularly apparent when the coverage in the Third List is high. So for example, when the Third List coverage is 0.9 and assuming that the Census and Survey coverage probabilities are 0.9 and 0.7, then there are 27 individuals expected to be counted in the (0,0,1)-cell. However, supposing there are an additional 10 erroneous people as a result of the Third List would imply that roughly a quarter of the individuals found in the (0,0,1)-cell count are not 'real'.

If on the contrary the Third List coverage is set at 30%, then there are 19 individuals expected to be found in the Third List only, the majority of them being erroneous. The effects on the estimators are clear: there will be an over-estimation of the population size. Accordingly, the simulations concentrated on the cases where the Census and Survey coverage probabilities both achieve moderate population coverage, i.e. 50% coverage of the population. It is of interest to determine if, given the Third List has coverage of 0.7, the four estimators give reasonable estimates of the population - supposing it is known that there are some erroneous enumerations in the observed cell counts.



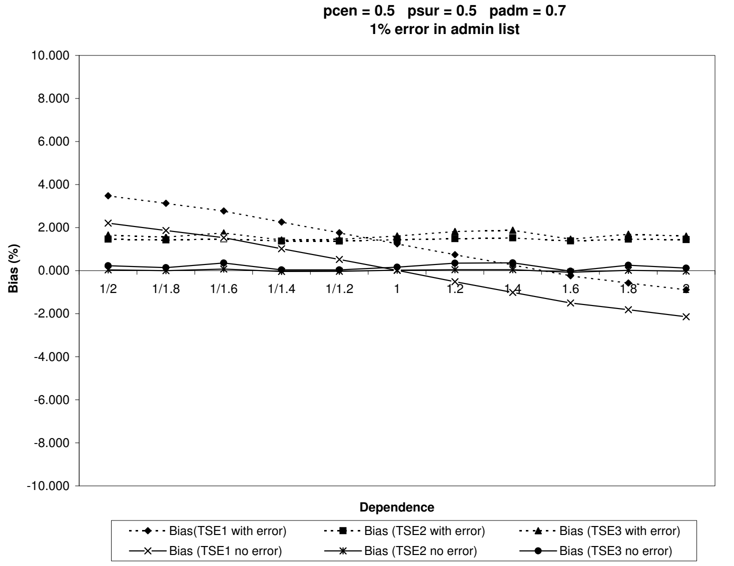


Figure 4.11: Bias of the triple and dual system estimators in the presence of erroneous enumerations. The bottom bottom plot just considers the three triple system estimators, while the top plot includes the dual system estimator.

Figure 4.11 shows what happens to the estimators in the presence of erroneous counts in the (0,0,1)-cell. It can be observed here that the population estimate under the dual system estimator remains unchanged since the presence (or for that matter, absence) of erroneous enumerations in the (0,0,1)-cell has no bearing on the calculation of the dual system estimate, as should be the case. Again as anticipated, the effect of introducing erroneous enumerations is to lead to an increase in the population size, but this increase is proportionate to the number of erroneous enumerations, due to the additivity property of log-linear modelling effects. As a demonstration, remembering that the estimate of N when the independence model holds (under certain conditions) can be re-written as

$\bar{\hat{N}} = \sqrt{\frac{n_{1++}n_{+1+}n_{++1}}{n_{111}}}$, it follows that when there are δ erroneous enumerations in the (0,0,1)-cell then the 'new' estimate of the population is

$$\bar{\hat{N}}^{*} = \sqrt{\frac{n_{1++}n_{+1+}\left(n_{++1} - \delta\right)}{n_{111}}} = \sqrt{\frac{n_{1++}n_{+1+}n_{++1}}{n_{111}} - \frac{n_{1++}n_{+1+}\delta}{n_{111}}} = \sqrt{\bar{\hat{N}}^{2} - \delta\frac{n_{1++}n_{+1+}}{n_{111}}}.$$

In the previous section, it was shown that $\bar{\hat{N}}$ is not the MLE, although it does exhibit similar unbiasedness when the coverage in the three lists is high. Also, under pairwise dependence between the Census and Survey, the estimate of the missing cell becomes,

$$\hat{n}^{*}_{000} = \frac{n_{110} + n_{100} + n_{010}}{n_{111} + n_{101} + n_{011}} \times (n_{001} - \delta) = \hat{n}_{000} - \delta\left(\frac{n_{110} + n_{100} + n_{010}}{n_{111} + n_{101} + n_{011}}\right).$$

This result follows from Table 3.6 in the section on the application of log-linear modelling to the census. This goes to show that supposing there are only erroneous enumerations in the Third List (i.e. in the (0,0,1)-cell) and an estimate of these can be found, say $\hat{\delta}$. Then, by simply extending the results in Chapter 3 it is possible to write down expressions of the missing cell that take into account the effect these erroneous enumerations have on the estimate $\hat{n}_{000}$.

4.3.2 Erroneous Enumerations in the Census and Third List

The simulations carried out above supposed that the erroneous enumerations could only be introduced through the Third List (i.e. the (0,0,1)-cell). The motivation was based on the fact that under dual system estimation as applied in the 2001 UK census, it was assumed that the erroneous enumerations had been removed through the matching processes, and as such the impact of overenumeration on the dual system estimates was considered to be minimal. The reality, however, is that there could be some erroneous enumerations in the Census counts, thereby influencing the dual system estimates.

So in the next part of the simulation study the estimators of the population were evaluated for their performance in the presence of some erroneous enumerations not only in the (0,0,1)-cell, but also in the (1,0,0) and (1,0,1)-cells. Nevertheless, it still remained important to keep the assumption that there were no erroneous enumerations in the Survey. In other words, there could be errors in the Census counts (as a result of its sheer operational size) and in the Administrative List (as a result of the complexities in assembling a comprehensive population register). But because the Survey is relatively small (in the 2001 UK census 300,000 households were sampled) it is reasonable to expect that there are stringent operational processes to prevent erroneous enumerations. It was of interest to also look at the behaviour of the estimators when there is an increase in the number of erroneous enumerations.

Therefore, in the following simulations it was assumed that there were now 50 erroneous people out of the simulated population of 1000 people and these erroneous counts could occur in any of the (0,0,1), (1,0,1) and (1,0,0)-cells. Figures 4.12 and 4.13 give the bias of the different estimators under two conditions - firstly when the Census and Survey achieve 50% coverage and the Administrative List achieves a coverage of 70% of the population, and secondly when the Survey and Administrative List achieves 70% coverage and the Census achieves 90% coverage.

Figure 4.13 shows that when there are erroneous enumerations in both the Census and Administrative List but the coverage rates are moderately high then the dual system estimator, although biased, is less biased than **all** the triple system estimators. The implications of this result are that if there are erroneous enumerations on the Census and the Third List, but the Census and Survey have high enough coverage then the dual system estimator seems to provide a better estimate of the population than the triple system estimators. However, it appears to confirm the assertion that the functionality of the Third List in triple system estimation is impaired when there are erroneous enumerations present. Hence, it seems that fitting the above dual and triple system estimators in the presence of erroneous enumerations leads to wrong estimates of the population - there needs to be an explicit adjustment of the population counts to take account of the erroneous counts. The introduction of a latent variable is the more obvious approach of doing this, where the latent variable here is the unobserved construct of a person's enumeration status (real or erroneous) that is imperfectly measured by the observed indicators of a person's enumeration by the Census, Survey or Third List.

Figure 4.12: Bias of the estimators in the presence of erroneous enumerations in both the Census and Third List (for p_{cen}=0.5 and p_{sur}=0.5, and p_{adm}=0.7).

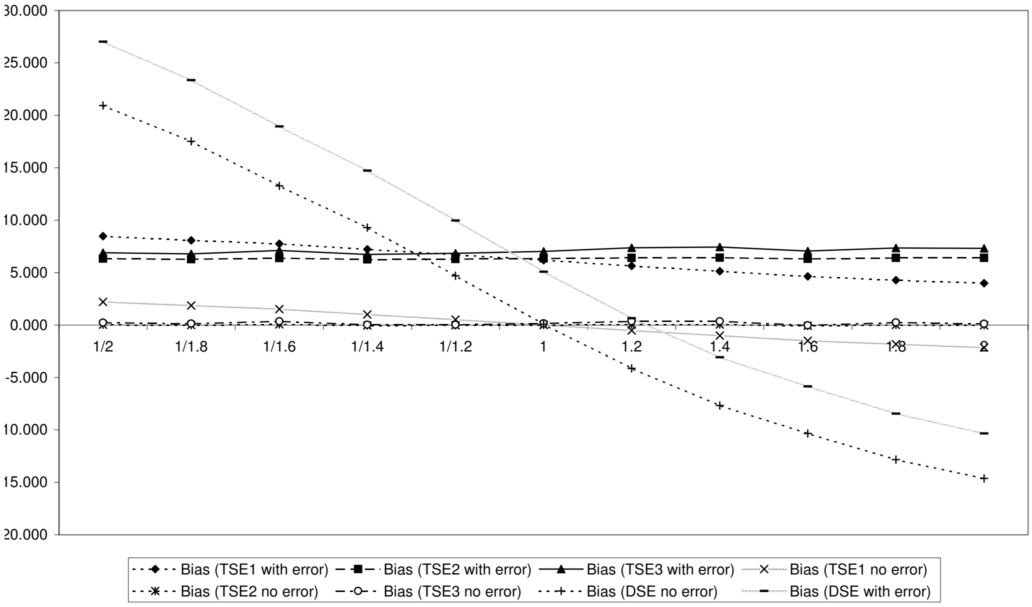
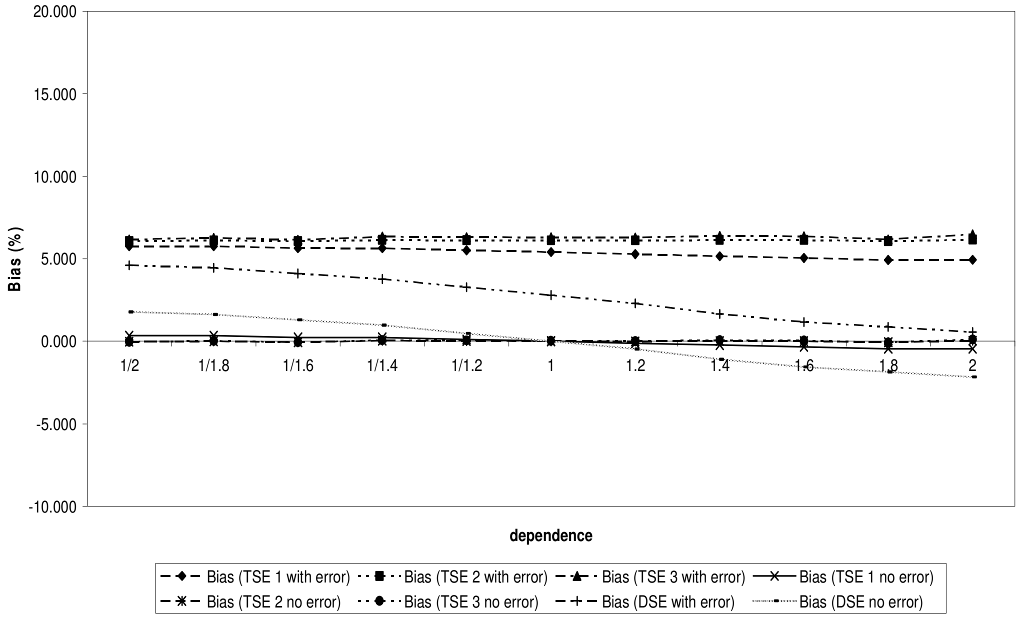


Figure 4.13: Bias of the estimators in the presence of erroneous enumerations in both the Census and Third List (for p_{cen}=0.9, p_{sur}=0.7 and p_{adm}=0.7).



4.4 Conclusion

It is true that independence, be it in dual system estimation or triple system estimation, is unlikely to hold in an actual census environment. This is because there is definite evidence of dependence between the Census and Survey (albeit there is less of a dependence between these two and the Third List). Nonetheless, the independence model which assumes no relationship between the different counts of the population does have some benefits. The single most important reason for its choice is that of model parsimony - the independence model is simplistic, and in most cases does approximate the true cell probabilities well, especially when the coverage probabilities are high. The mean squared error of the independence model is lower than that of the partial dependence and 'saturated' models, for all the simulations considered. This is because, although the independence model is the most biased, it also has the smallest variance as it is based on estimating fewer parameters. In essence, the mean squared error is smaller because the bias does not dominate the variance. In the same vein, although the 'saturated' model gives unbiased estimates of the population, it has been shown to be inefficient when the data has been simulated under the Census and Survey pairwise dependence model.

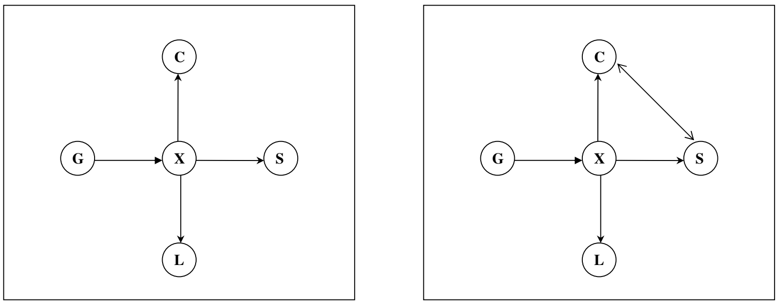
The motivation of the simulation work was the desire to have some idea as to what constitutes 'low' and 'high' levels of coverage and consider if there is the need for triple system estimation. The simulations considered permutations of four probabilities of 0.3, 0.5, 0.7 and 0.9 for the Census, Survey and Third List coverage. It was found that if the

Census manages to enumerate roughly 90% of the population and the Survey achieves 70% then the dual system estimate is fairly unbiased, and bringing in a Third List into frame does not significantly improve upon the population estimates. However, it has been shown that in the presence of dependence, and for a Census or Survey that only counts 50% of the population, there are definite advantages of using data from an administrative source to obtain population size estimates that have been adjusted for underenumeration. Further, the simple triple system estimator that assumes independence between the Census, Survey and Third List is found to be very efficient, even in the presence of some dependence.

In the investigation of the erroneous enumerations, the fairly reasonable assumption was made that the computer and probability matching procedures are sufficiently advanced to remove any duplicates in the Census or Survey. As such, it was presumed that any erroneous enumerations will result only from the Third List being brought into the fray. Though this assumption may be considered too simplistic, as it implies that the erroneous enumerations only exist in the cell count corresponding to those missed by the Census and Survey but counted by the Third List, it does have some justification. In the 2001 census, it was assumed that for individuals counted in the selected Census Coverage Survey postcodes it was possible to ascertain whether they were counted in the Census only, Survey only or both. If the Third List is matched to the Census and Survey (which have no erroneous enumerations, by definition), then for the sampled postcodes it follows that the only way erroneous enumerations can be introduced is via the (0,0,1)-cell. What the simulations showed was that this leads to an over-estimation of the population, by a factor proportional to the number of erroneous enumerations in the (0,0,1)-cell. In actuality, however, there will be some erroneous enumerations added through the Census process because of the scale of the operation, although it may still be reasonable to assume that there are none introduced through the Survey process. So there was an investigation into the effect of erroneous people in the (1,0,1), (1,0,0) and (0,0,1)-cells. When this happens the simulations showed that the triple system estimators become more biased and supposing the Census and Survey achieved moderately high levels of coverage of the population it turned out that the dual system estimator fared better than the triple system estimators.

The proposed way of dealing with erroneous enumerations is to use latent class analysis by assuming that the Census, Survey and Third List are imperfect indicators of an individual's true enumeration status - which cannot be directly observed and hence a latent variable. In the basic latent class model the latent variable is deemed to be locally independent of the observable variables, which substantively means that the associations observed amongst the Census, Survey and Third List are only attributable to the each of their relationships to the latent variable. When there is some additional variation that is left unaccounted for by the latent variable, then a locally dependent model is needed to be fitted to the data. In a triple system context this local dependent latent class model is non-identified. To fit a locally dependent model that is identified when the method adopted in this thesis is to introduce a grouping covariate. There are, however, a number of conditions under which such a model within a capture-recapture context with an unobserved cell can be fitted. Firstly, the observed cells should be sufficient statistics for the missing cell, so that the EM algorithm can be utilized to find a suitable estimate under a postulated unsaturated model. Secondly, the grouping covariates should be associated to the latent variable only, in other words the relationships between these grouping covariates and the manifest variables - i.e. Census, Survey and Third List - are completely mediated through the latent variable (as illustrated in Figures 4.14(a) and 4.14(b)). It will be shown in the next chapters that even though this model is identified and as such the population size estimate that makes adjustments for both underenumeration and overenumeration can be obtained, in general the standard errors of these estimates can be difficult to derive.

Figure 4.14: Latent Class models with a covariate effect, G



(a) Local Independence

(b) Local Dependence

Chapter 5

An Application of Log-linear Modelling to Census Underenumeration

5.1 Introduction

Dual system estimation relies on the assumption of list independence. However, as the previous chapters have shown, the initial census and follow-up survey have been widely viewed as being related. With only two samples, it is not possible to investigate the extent of this dependence, but when there is another list the log-linear modelling framework can be used. The previous chapters also highlighted the problem of differential underenumeration. Therefore, in response to the problem of differential underenumeration, especially when the follow-up survey achieves poor coverage of the population, the US Census Bureau carried out a triple system exercise during the Dress Rehearsal leading up to the 1990 Census. At the heart of the matter was that, after the 1980 US decennial census there was the realisation that there was differential undercount of Blacks and other ethnic minorities (see Ericksen et al. (1985) and Wolter (1990)).

Accordingly in the run-up to the 1990 US Census, the Census Bureau compiled an administrative records list called the Administrative List Supplement (ALS) for their 1988 Census Dress Rehearsal. As aforementioned, the purpose of the ALS was to provide a check on the 1990 post enumeration survey methodology. This ensured that as part of the Dress Rehearsal data from the Census, the post-enumeration survey and the ALS was gathered for some sampled blocks. This, therefore, made it possible for the US Census Bureau to trial out the feasibility of the triple system methodology in a ‘real’ census environment. Zaslavsky and Wolfgang (1990, 1993) were able to show that the third list can be used to investigate list dependence, provide more precise estimates of the hard-to-count population and improve the overall census coverage.

The data set considered in this chapter is from a population subgroup from the 1988 Dress Rehearsal carried out in St Louis, Missouri. On the basis of the 1980 US Census, it was found that the level of undercount was greatest for Black males. Actually, the national undercount rate for Blacks has remained roughly 5% higher than for Non-blacks in every census since 1940, and the group believed to be most seriously undercounted in censuses are Black male renters (Darroch et al. (1993)). Also it was established that the same factors that had an effect on whether or not a person was missed in the census had an influence on the post-enumeration survey. This meant that the most undercovered population group, i.e. Black male adults, also had the tendency to be underenumerated by the survey (Wolfgang (1989); Zaslavsky and Wolfgang (1990)). To that end, the sampling frame for the Dress Rehearsal was restricted to select participants on the basis of age, sex and race to fall within four post-strata: Black Males aged 20-29 in Owned homes (O2), Black Males aged 30-44 in Owned homes (O3), Black Males aged 20-29 in Rented homes (R2) and Black Males aged 30-44 in Rented homes (R3). Furthermore, the ALS was specifically designed and assembled from government rosters that targeted Black male living in rented accommodation. It is clear from the design of the Dress Rehearsal that the simplistic dual system estimator - or, for that matter, the triple system estimator - that assumes list independence will be expected to yield biased results.

The Dress Rehearsal encountered some problems due to matching and classification issues that were unresolved which meant that the resulting data were not identical to the raw data obtained from the three systems used to capture the population. Unlike in animal capture experiments where tags and marks are used to identify the animals across captures, it was found to be very difficult to identify people who appeared in the different sources. The method used in the Dress Rehearsal was to match across the three sources based on some key demographic variables of age, name and sex. But it was difficult to determine correct matches, which implied that a large number of cases were removed and this meant that the final 2x2x2 matched data set had just over 1,000 observed people. Nevertheless, it was possible to obtain data that classified the respondents into whether or not they appeared on the Census, Post-enumeration Survey or the Administrative List Supplement, post-stratified by age and tenure.

The data from the 1990 US Census seem dated, and it would have been ideal to use data from the 2000 Census. But in 2000, following on from the 1999 US Supreme Court judgement that ruled that sampling for non-response follow-up was not consistent with the US Census Act for providing apportionment counts for legislative representation, the focus unfortunately turned away the production of triple system counts in the format of the 1990 Census Dress Rehearsal. The Administrative Records Experiment (AREX) which was carried out in 2000 (in selected areas of the US - Baltimore City, Baltimore County and three counties in Colorado) sought to investigate the feasibility of using an administrative records census to replace the traditional census. It did, however, also investigate how information from administrative records can augment the Master Address

File (the US equivalent of the UK Postal Address File) in pre-Census operations, and to use administrative information to directly substitute for any missingness post-Census (see Cohen and King (2000)). Nevertheless, although there are some legalities involved that prevent the implementation of triple system estimation (in the format proposed in this thesis) in the US, there has been a great deal of theoretical research focused on the use of administrative records in the census (Judson (2000), Biemer et al. (2001a), Stuart and Zaslavsky (2002), for example).

5.2 Triple System Results for Data from the 1988 Dress Rehearsal Census in St Louis

Table 5.1: Triple System Estimation data from 1988 Dress Rehearsal

Cell	O2	R2	O3	R3
n_{000}	-	-	-	-
n_{001}	59	43	35	43
n_{010}	8	34	10	24
n_{011}	19	11	10	13
n_{100}	31	41	62	32
n_{101}	19	12	13	7
n_{110}	13	69	36	69
n_{111}	79	58	91	72
n	228	268	257	260

The counts of respondents in the Dress Rehearsal for the four post-strata are given in Table 5.1. One thing to notice is that the third list (i.e. the ALS) has reasonably decent coverage of the population of interest. Apart from owners in the 30-44 age group - stratum O3 - there are more people found in only the ALS (i.e. (0,0,1)-cell) than only in the Census (i.e. (1,0,0)-cell). This does support the presumptions of the US Census Bureau in carrying out the Dress Rehearsal that the census processes fail at obtaining a good enough coverage of black male renters. It also does not come as a surprise that the only post-stratum that the Census does better than the ALS is O3, the older house owners. Furthermore in the rented categories, there are fewer people found by the Survey alone, i.e. the (0,1,0)-cell. Although it must be said that for renters aged 20-29, R2, the post-enumeration survey works well. This may be due to the fact that the survey processes were specifically designed to capture rented households with young Black male residents.

Initially it is good to explore if there is some evidence to suggest whether dual system estimation is appropriate here. The DSE in this case can be obtained by summing over one of the three systems, for instance summing over the administrative list yields

$$\frac{n_{10+}n_{01+}}{n_{11+}n_{00+}} = 1, \quad \text{and hence} \quad \hat{n}_{000} = \frac{n_{10+}n_{01+}}{n_{11+}} - n_{001}.$$

Since there are three sources, there are three possible ways of calculating the DSE, and these have all being considered. Table 5.2 gives the results of the estimates of the missing under the different DSEs. Albeit, realistically, the only one of interest is the first estimate (DSE1) based on the Census and Survey, and summing over the administrative list information.

Table 5.2: Estimates of the missing cell count $\hat{n}_{000}$ under dual system estimation

	O2	R2	O3	R3
DSE 1 (Census and Survey)	-44.33	-24.22	-23.19	-32.77
DSE 2 (Census and Admin List)	27.02	50.85	32.40	47.59
DSE 3 (Survey and Admin List)	-14.22	41.10	-40.14	22.71

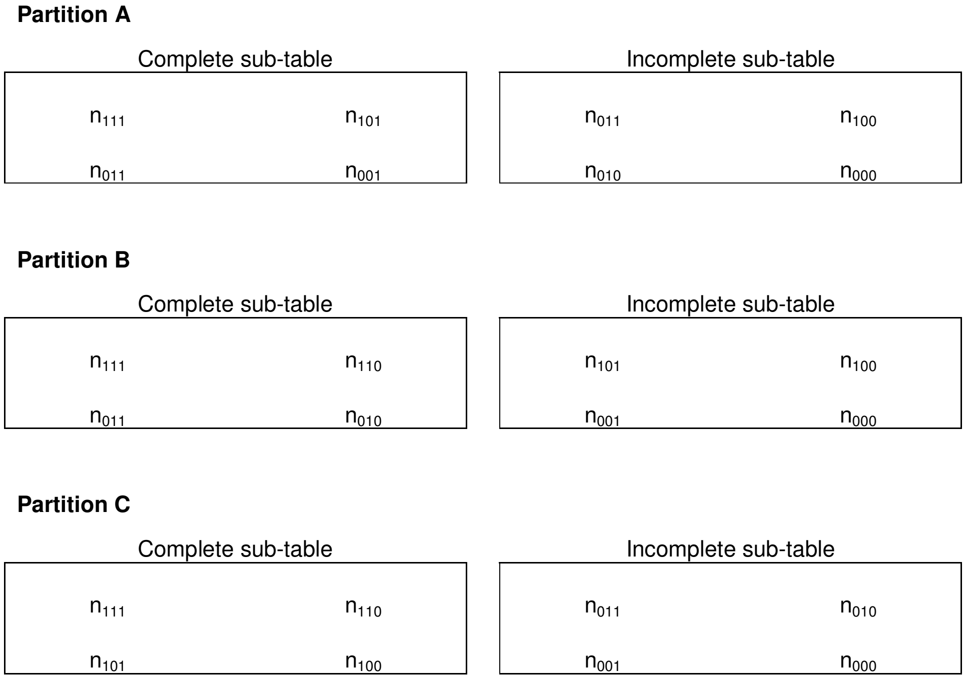
The dual system estimate calculated marginal over the administrative list gives a negative estimate of the missing cell, $\hat{n}_{000}$, for all the post-strata. This means that the dual system estimate $\hat{n}_{00+}$ is less than the observed individuals added by the administrative list, i.e. persons in the (0,0,1)-cell.

There is therefore a substantial under-estimation of the population because coverage in the Census and Survey is very low. It can be anticipated that the administrative list coverage is also low, thus there will be additional people missed by all three sources. Furthermore, it may be of interest to note that the missing cell estimate computed after summing over the Survey (i.e. DSE2) is positive for all four post-strata. One reason could be because the Survey coverage amongst the population of interest - i.e. Black males - was very poor, but the ALS went some way at compensating for this (though, as will be shown, not necessarily in an independent manner). This point is further illustrated by the positive estimates of DSE3 for respondents living in rented households (R2 and R3) - in effect, this time the ALS is compensating for the low coverage in the Census.

There was a preliminary investigation that focused on the odds ratios in the 2x2 subtables. The results are indicative of the strength of departure from list independence. Figure 5.1 shows how the odds ratios were calculated. It was mentioned previously that the 2x2x2 contingency table can be partitioned into two 2x2 subtables, with a complete subtable and an incomplete subtable. The partitioning can be done in three ways - controlling for being observed in the Administrative List (Partition A), being observed in the Survey (Partition B) or being observed in the Census (Partition C). The homogeneity assumption implies that the mechanism that underlies a person being counted in the Census should be similar to that of a person being missed in the Census. Thus the odds ratio in the complete subtable, obtained when controlling for Census enumeration, can be indicative of the odds ratio in the incomplete subtable, under this assumption. Similar assertions can be made when controlling for Survey enumeration and Administrative List enumeration.

If there is independence of two of the systems given the third, then it is expected that the odds ratios in Table 5.3 should be close to 1. Therefore, if these odds are substantially different from 1 it becomes understandable to question the independence assumption. This result supports the results presented in Table 5.2, which show that DSE2 (i.e. the estimate found under Partition B) is the one that consistently gives positive estimates of the missing cell count, $\hat{n}_{000}$ - admittedly this is still an under-estimate. Zaslavsky and Wolfgang (1993) used the jack-knife to calculate the standard errors of log-odds ratios. They found that although the standard errors were large, the log-odds ratios were at least three times the standard errors. They also observe that the odds ratios under Partition B and C are closer to 1, than Partition A. This suggests that the Administrative List is more nearly independent of the Census or Survey. Clearly the Census and Survey have much more similar data collection methods than the ALS, so this result does make sense on consideration.

Figure 5.1: Partitioning of the 2x2x2 contingency table.



After the preliminary analysis of the odds ratios, some log-linear models are subsequently fitted to the data in order to estimate the missing cell. The estimates of the standard errors will be obtained using the Supplemented EM (SEM) algorithm. The odds ratios in Table 5.3 show that there is some degree of dependence between the sources, and so a simple dual system estimate will under-estimate the size of the missing population.

Table 5.3: Odds ratios for the complete sub-tables

	Odds Ratio		
	Partition A (ALS)	Partition B (Survey)	Partition C (Census)
O2	12.91	2.56	9.92
R2	18.89	2.60	2.87
O3	24.50	2.53	12.06
R3	34.02	1.93	4.77

A program called **EM.sim** was written in SPLUS/R that fitted the different log-linear models to the data. It uses the EM algorithm to find the missing cell estimate $\hat{n}_{000}$ that keeps the posited relationship of the observed cells $\{n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}\}$. It relies on the *glm* function to fit the log-linear model in the M-step, and estimate $\hat{n}_{000}$ in the E-step. The program can be found in Appendix B.1.

The program first starts by assuming that the estimate of the missing cell is zero, i.e. there are no individuals missed in all three lists. Based on the log-linear model fitted to the seven observed counts, it checks whether a value of zero is a suitable estimate. If this is not true, a new estimate of the missing cell is fitted, with the iterative process continuing until the difference between iterative fitted estimates are close enough to a pre-determined convergence criterion.

In all cases, there is evidence from the observed cells to discount the suggestion that the combined coverage on the three lists is good enough to render the missing cell count negligible. Therefore, a model (or some other method) needs to be fitted to estimate this cell. Table 5.4 gives the estimate of the missing cell $\hat{n}_{000}$ and the goodness of fit statistic for each of the models (see Table 3.6). In addition, the program computes two goodness of fit measures that can be used to assess how well any of the eight models is consistent with the observed cell counts $\{n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}\}$. These are the log-likelihood chi-squared statistic G^2 and the Pearson chi-squared statistic X^2. Another program (given in Appendix B.2) was written to implement the SEM algorithm to produce the asymptotic covariance matrix of the estimated log-linear parameters. The square roots of the diagonal elements are the asymptotic standard errors. These SEM variances were compared to those calculated using the parametric bootstrap (see SPLUS/R program in Appendix B.5) and also those derived under the Delta method.

From Table 5.4, the three sources have some definite inter-relationships, and the size of the likelihood statistics show that the model assuming complete independence poorly fits the data. Table 5.5 gives the p-values for the different log-linear models using the likelihood ratio statistic as this represents the deviance. Here the best model is the one for which the deviance does not exceed the critical value for the appropriate number of degrees of freedom.

Therefore, there is some evidence (as exhibited in Table 5.3) to suggest that the best fitting model is the one that accounts for the pairwise interactions between the Census and Survey and the Survey and Administrative List. In other words, the Census and Administrative List are conditionally independent of each other, given the Survey. This is an intuitively reasonable model considering that there are different enumeration processes underlying the Census or the ALS. In fact the ALS was specifically designed to find people who were hard-to-count in the Survey; also the Census and Survey were not operationally independent as expected.

Table 5.4: Estimate of the missing cell count and likelihood statistics (X^2 and G^2) under different models

Model		O2	R2	O3	R3	df
Independence	$\hat{n}_{000}$	13.78	28.43	14.32	18.21	3
	G^2	72.59	54.83	90.19	76.20	
	X^2	68.68	54.31	83.48	76.39	
{L,CS}	$\hat{n}_{000}$	24.02	25.96	24.35	17.30	2
	G^2	59.01	54.23	62.54	76.06	
	X^2	56.71	52.58	65.05	75.04	
{S,CL}	$\hat{n}_{000}$	7.86	23.65	8.03	12.78	2
	G^2	68.55	52.80	84.54	70.73	
	X^2	69.59	50.40	77.26	67.42	
{C,SL}	$\hat{n}_{000}$	26.22	76.43	33.16	58.42	2
	G^2	34.46	12.19	59.27	15.71	
	X^2	34.87	11.87	55.71	14.68	
{CS, CL}	$\hat{n}_{000}$	19.07	20.20	17.20	11.13	1
	G^2	58.71	51.58	61.25	69.99	
	X^2	55.69	48.23	61.29	64.84	
{CS, SL}	$\hat{n}_{000}$	96.22	146.78	166.77	196.23	1
	G^2	3.15	6.53	3.55	3.04	
	X^2	3.45	6.23	3.77	2.98	
{CL, SL}	$\hat{n}_{000}$	24.84	132.79	34.99	79.34	1
	G^2	34.44	8.78	59.25	14.73	
	X^2	34.71	8.34	55.70	13.60	
‘Saturated’	$\hat{n}_{000}$	245.11	379.69	418.83	378.68	0
	G^2=X^2	(0)	(0)	(0)	(0)	

The best fitting model can be represented by the log-linear model

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)}. \tag{5.1}$$

This model, given in equation (5.1), was fitted to the different post-stratified tables O2, R2, O3 and R3. However, the advantage of the log-linear modelling framework is that it can be extended to include the post-stratified variables as covariates in the model. The post-stratified variables can be thought of as a grouping covariate G, such that equation (5.1) becomes

$$\log \mu_{ijkg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_g^{(G)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)} + \lambda_{ig}^{(CG)} + \lambda_{jg}^{(SG)} + \lambda_{kg}^{(LG)} + \lambda_{ijg}^{(CSG)} + \lambda_{jkg}^{(SLG)}. \tag{5.2}$$

It follows that G has four levels, namely Young Owners, Young Renters, Old Owners and Old Renters. The program **EM.sim** has the capability to fit the log-linear model specified by equation (5.2) to the data, the results of this are presented below. Before fitting the model in SPLUS/R, the data needs to be re-formatted as the data frame in Table 5.6.

Table 5.5: Goodness of fit of the models using the G^2 statistic

Model		O2	R2	O3	R3
Independence	G^2	72.59	54.83	90.19	76.20
	p−value	<0.0001	<0.0001	<0.0001	<0.0001
{CS}	G^2	59.01	54.23	62.54	76.06
	p−value	<0.0001	<0.0001	<0.0001	<0.0001
{CL}	G^2	68.55	52.80	84.54	70.73
	p−value	<0.0001	<0.0001	<0.0001	<0.0001
{SL}	G^2	34.87	11.87	55.71	14.68
	p−value	<0.0001	0.0026	0.0001	0.0006
{CS,CL}	G^2	58.71	51.58	61.25	69.99
	p−value	<0.0001	<0.0001	<0.0001	<0.0001
{CS,SL}	G^2	3.15	6.53	3.55	3.04
	p−value	0.0759	0.0106	0.0595	0.0812
{CL,SL}	G^2	34.44	8.78	59.25	14.73
	p−value	<0.0001	0.003	<0.0001	0.0001

Table 5.6: SPLUS/R Data frame for the log-linear model with the post-strata included

Cell	Count	Cell	Count
O2		O3	
n_{0001}	-	n_{0003}	-
n_{1001}	31	n_{1003}	62
n_{0101}	8	n_{0103}	10
n_{1101}	13	n_{1103}	36
n_{0011}	59	n_{0013}	35
n_{1011}	19	n_{1013}	13
n_{0111}	19	n_{0113}	10
n_{1111}	79	n_{1113}	91
R2		R3	
n_{0002}	-	n_{0004}	-
n_{1002}	41	n_{1004}	32
n_{0102}	34	n_{0104}	24
n_{1102}	69	n_{1104}	69
n_{0012}	43	n_{0014}	43
n_{1012}	12	n_{1014}	7
n_{0112}	11	n_{0114}	13
n_{1112}	58	n_{1114}	72

Table 5.7: Estimates of the missing population in each group

	Young Owners	Young Renters	Old Owners	Old Renters
Estimate	96.26	146.92	166.92	196.57

Table 5.7 gives the estimates of the missing cells under the model with covariate effects, given in equation (5.2). The table shows that the estimates of the missing cell are the

same under equations (5.1) and (5.2). The Pearson and likelihood ratio statistics for this model are 16.43 and 16.27 respectively. Since the likelihood ratio statistic that tests the null hypothesis that the model holds against the saturated model has the property of being additive, it follows that the G^2 statistic under model (5.2) is the sum of the four G^2 statistics for the model (5.1) fitted to O2, R2, O3 and R3, i.e. $16.27 = 3.15 + 6.53 + 3.55 + 3.04$.

Note that there is an advantage of comparing the likelihood statistics this way under models (5.1) and (5.2) in that it can highlight particular aspects of model failure. It can be seen that the G^2 statistic of 6.53 for R2 is almost 2 times the other group statistics. In the model selection process other models were considered; for example, removing the $\lambda_{ijg}^{(CSG)}$ and $\lambda_{jkg}^{(SLG)}$ terms made the model less complicated, but this proved to be at the detriment of the model fit. However, this could be an advantage sometimes, in particular when there are model identifiability issues.

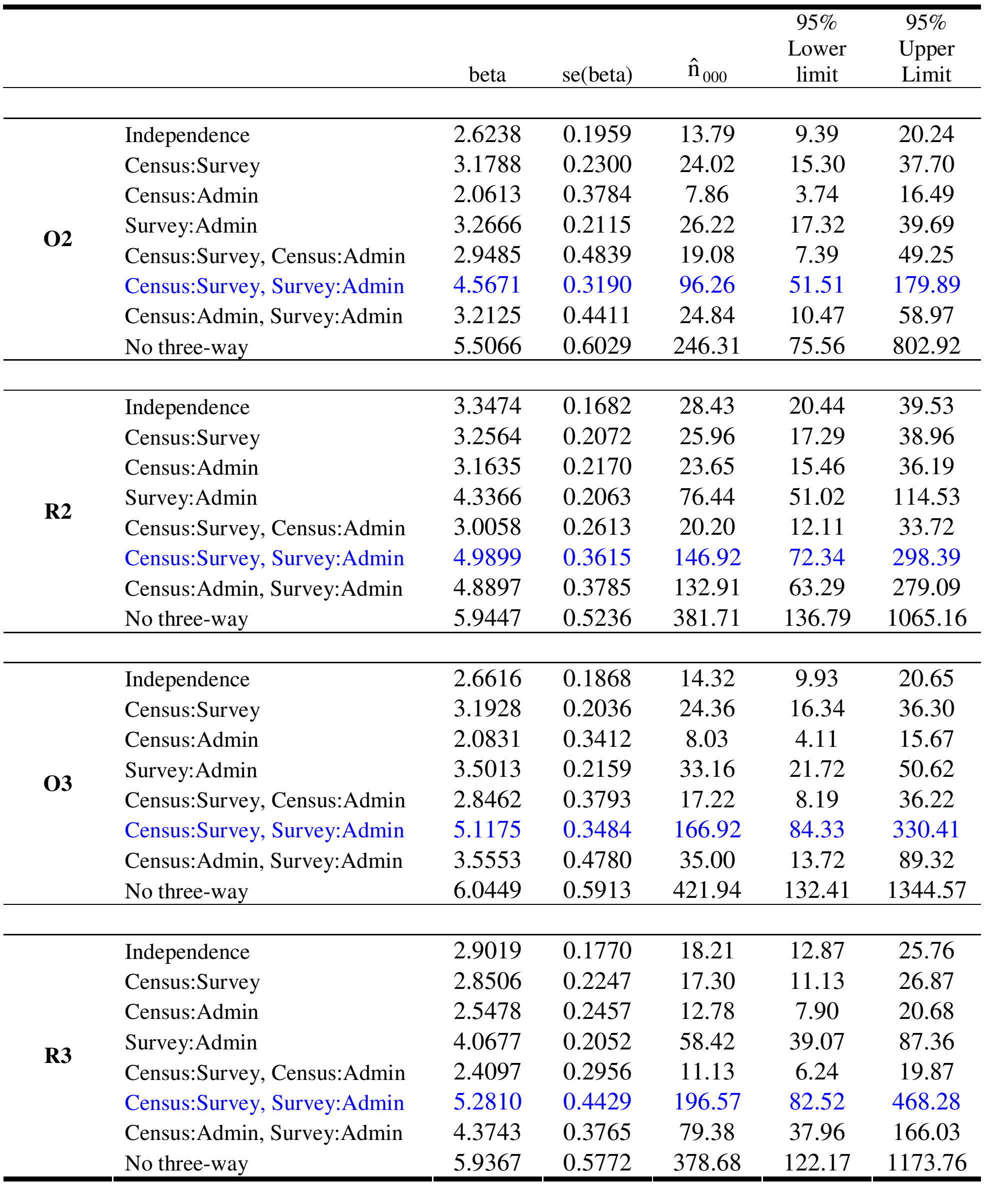
There is an additional advantage of fitting the log-linear model with covariates in that it converges relatively quicker. **EM.sim** takes 75 steps to fit the first model to the O2 table, 126 steps to fit the R2 table, 133 for the O3 table, and 222 iterations for the R3 table. However, to find all the four estimates, **EM.sim** converges in 171 steps.

There is further point to make, which happens to be perfectly highlighted by the US Census data. It may be noticed that the estimate of the missing cell under the 'selected' parsimonious model and the 'saturated' homogeneous association model are seemingly different. The four estimates of the missing in the groups are 96.22, 146.78, 166.77 and 196.23, under the best model, while the estimates using the 'saturated model' are 245.11, 379.69, 418.83 and 378.68, respectively. It is cause for concern that these two models give such different estimates, and both are seemingly correct; the only way to choose between them will be to know the truth, which is unfortunately not possible.

There is another explanation that could explain why these seemingly different estimates result from similar models. Consider the confidence intervals of the estimates of the selected models (given in Table 5.8, overleaf). For post-strata R2, O2 and R3 the confidence intervals for the best selected model are respectively (72.34, 298.39), (84.33, 330.41) and (82.52, 468.28). These confidence intervals all contain the estimates under the 'saturated' model, which implies estimates of 379.69, 418.83 and 378.68 are perfectly feasible under the parsimonious model. However, for post-stratum O2 the estimate under the saturated model of 245.11 lies outside the confidence interval of (51.51, 179.89), which could be indicative of model failure here. An alternative interpretation will be presented in Chapter 6.

5.3 Variance Estimation

Table 5.8: Estimates of standard errors of model parameters with appropriate confidence intervals (using the SEM algorithm)



		beta	se(beta)	$\hat{n}_{000}$	95% Lower limit	95% Upper Limit
O2	Independence	2.6238	0.1959	13.79	9.39	20.24
	Census:Survey	3.1788	0.2300	24.02	15.30	37.70
	Census:Admin	2.0613	0.3784	7.86	3.74	16.49
	Survey:Admin	3.2666	0.2115	26.22	17.32	39.69
	Census:Survey, Census:Admin	2.9485	0.4839	19.08	7.39	49.25
	Census:Survey, Survey:Admin	4.5671	0.3190	96.26	51.51	179.89
	Census:Admin, Survey:Admin	3.2125	0.4411	24.84	10.47	58.97
	No three-way	5.5066	0.6029	246.31	75.56	802.92
R2	Independence	3.3474	0.1682	28.43	20.44	39.53
	Census:Survey	3.2564	0.2072	25.96	17.29	38.96
	Census:Admin	3.1635	0.2170	23.65	15.46	36.19
	Survey:Admin	4.3366	0.2063	76.44	51.02	114.53
	Census:Survey, Census:Admin	3.0058	0.2613	20.20	12.11	33.72
	Census:Survey, Survey:Admin	4.9899	0.3615	146.92	72.34	298.39
	Census:Admin, Survey:Admin	4.8897	0.3785	132.91	63.29	279.09
	No three-way	5.9447	0.5236	381.71	136.79	1065.16
O3	Independence	2.6616	0.1868	14.32	9.93	20.65
	Census:Survey	3.1928	0.2036	24.36	16.34	36.30
	Census:Admin	2.0831	0.3412	8.03	4.11	15.67
	Survey:Admin	3.5013	0.2159	33.16	21.72	50.62
	Census:Survey, Census:Admin	2.8462	0.3793	17.22	8.19	36.22
	Census:Survey, Survey:Admin	5.1175	0.3484	166.92	84.33	330.41
	Census:Admin, Survey:Admin	3.5553	0.4780	35.00	13.72	89.32
	No three-way	6.0449	0.5913	421.94	132.41	1344.57
R3	Independence	2.9019	0.1770	18.21	12.87	25.76
	Census:Survey	2.8506	0.2247	17.30	11.13	26.87
	Census:Admin	2.5478	0.2457	12.78	7.90	20.68
	Survey:Admin	4.0677	0.2052	58.42	39.07	87.36
	Census:Survey, Census:Admin	2.4097	0.2956	11.13	6.24	19.87
	Census:Survey, Survey:Admin	5.2810	0.4429	196.57	82.52	468.28
	Census:Admin, Survey:Admin	4.3743	0.3765	79.38	37.96	166.03
	No three-way	5.9367	0.5772	378.68	122.17	1173.76

Table 5.9: Estimates of standard errors using the grouped data (using the SEM algorithm)

	beta	SE(beta)	$\hat{n}_{000}$	95% Lower Limit	95% Upper Limit
O2	4.5671	0.3191	96.26	51.50	179.93
R2	4.9899	0.3619	146.92	72.28	298.62
O3	5.1175	0.3496	166.92	84.12	330.21
R3	5.2810	0.4429	196.57	82.51	468.29

Having managed to obtain the estimates of the missing cell under different models, it now remains to provide some estimates of the precision by computing the standard errors, and the respective confidence intervals. Now, the EM algorithm used to obtain the above estimates of the missing cell does so by maximizing the incomplete likelihood. However, the information matrix in the observed data for the model under the EM algorithm is complicated as it requires differentiating and inverting this complicated likelihood, which can be computationally unstable ((Little and Rubin, 2002, page 191). As such, the supplemented EM (SEM) algorithm has been proposed to get around this problem of instability of the information matrix by using the expected complete data information and a matrix defined by the rate of convergence of the EM algorithm. Accordingly, since the EM algorithm has been used to find estimates of the missing, the SEM algorithm is used to determine the variances of the model estimates. Obviously one of the recurring arguments in capture-recapture methods is on whether the assumption of normality is necessarily valid, so the Delta method asymptotic variances and confidence intervals were compared to those found under the SEM algorithm and the Bootstrap.

The SEM was implemented for the US Census data, the results of which appear in Tables 5.8 and 5.9. Table 5.8 gives the standard errors and confidence intervals for each of the eight hierarchical models. For Table 5.9, the SEM algorithm was implemented to the data with covariate information (as shown in Table 5.6). Both yield pretty similar results, and the differences can be attributed to rounding errors at different stages of the computation. Also it can be observed that the 95% confidence intervals are skewed, with the missing cell maximum likelihood estimate being nearer to the lower end of the confidence interval.

It may be noted that both the Delta method and SEM algorithm produce asymptotic variances. The Delta method produces confidence intervals for N and assumes that $\hat{N}$ is asymptotically normal such that the confidence interval is centred around $\hat{N}$. Unfortunately, the distribution of $\hat{N}$ is skewed in practice and so the above confidence interval can give misleading results (Coull and Agresti (1999)). However, under the SEM algorithm because the estimation is carried out on a different scale, the confidence intervals produced are skewed. Furthermore, it is common for $\hat{N}$ to be nearer to the lower end of the interval (Van Deusen (2002)). Thus it follows that the SEM variances might be expected to produce more realistic confidence intervals than the Delta method. The bootstrap, in contrast, may be more robust to data that exhibits skew, since though it is computationally intensive the basic ideas of the bootstrap do not rely on any distributional assumptions. The results of the bootstrap standard errors are given in Table 5.10 and it can be seen that they are similar to the SEM standard errors in Tables 5.8 and 5.9.

Table 5.10: Estimates of the empirical standard errors of the model parameters with appropriate confidence intervals (using the Bootstrap)

	beta	SE(beta)	$\hat{n}_{000}$	95% Lower Limit	95% Upper Limit
O2	4.5675	0.3252	96.31	51.91	182.16
R2	4.9909	0.3448	147.07	74.82	289.08
O3	5.1156	0.3747	166.60	79.93	347.24
R3	5.2835	0.4343	197.06	84.12	461.61

These SEM and Bootstrap confidence intervals are different to those derived under the Delta method. It may be recalled that in Section 3.7 the results of the asymptotic variance of the population total $\hat{N}$ using the Delta method are provided for different log-linear models. The formula for the selected best fitting model, with pairwise relationships between the Census and Survey and Survey and Administrative List, is given by

$$\hat{V}\left(\hat{N}\right) = \left(\hat{n}_{000}\right)^2 \left[\frac{1}{n_{101}} + \frac{1}{n_{001}} + \frac{1}{n_{100}} + \frac{n_{101}}{n_{001}n_{100}}\right]. \tag{5.3}$$

For the US Census data, the above formula, equation (5.3), was used to calculate the Delta method asymptotic variance and the confidence intervals of the population estimate, the results of which are displayed in Table 5.11.

Table 5.11: Estimates of the asymptotic standard errors of $\hat{N}$

	$\hat{N}$	Asymptotic SE	95% Lower Limit	95% Upper Limit
O2	324.26	32.25	261.05	387.47
R2	414.92	54.29	308.51	521.33
O3	423.92	59.63	307.05	504.79
R3	456.57	88.45	283.21	629.93

The 95% confidence intervals for the estimated population totals using the SEM method are (279.50, 407.92) for O2, (340.28, 566.62) for R2, (341.12, 587.21) for O3 and finally for R3 (342.51, 728.29). The corresponding bootstrap confidence intervals of the estimated population totals are (279.91, 410.16) for O2, (342.82, 557.08) for R2, (336.93, 604.24) for O3 and (344.12, 721.61) for R3. It can be recognized that these SEM and bootstrap intervals are wider and more skewed in comparison with the Delta method intervals.

5.4 Conclusion

It can be said with fair confidence that there is some value of an administrative list within census enumeration methodology. The above exercise using data - which admittedly was not of the best quality - has shown how possible it is to demonstrate the usefulness of administrative lists in investigating the previously untestable assumption of independence between the Census and Survey. So even though the Third List brings with it additional complications in that it was related to the Survey, the log-linear modelling framework is able to provide models that can account for sources of dependence in the estimation of the population size. One thing that is important is that the methodology used to assemble the Third list plays a vital role in ascertaining the dependence structure needed in the estimation of the missing cell, and ought to be factored into the estimation process.

However, the US Census application has highlighted problems with the no three-way interaction assumption. Under this assumption, which is needed in order to be able to estimate the missing cells, the implication is that although every variable may interact with each other variable, there is no interaction between the three variables (here, the Census, Survey and Third List interaction term is zero). However, provided there is *no unaccounted heterogeneity*, the model that includes the three-way interaction suggests that apparently all other models fail to represent the data in a suitable manner, which under model parsimony may be hard to believe. The belief is that the no three-way interaction model is the closest to the saturated model and furthermore even this model is anticipated to over-fit the observed data. Thus it is now possible to posit other models with fewer interaction terms that could better fit the observed data. Nonetheless, when there is some doubt regarding this assumption understandably the population size estimates obtained may be incorrect. In the next chapter, it will be shown that by extending the log-linear model to include a latent variable there is an improvement.

Chapter 6

Estimation of Population Totals from Imperfect Data

6.1 Introduction

This chapter aims to expand on the methodology developed thus far in order to estimate the population size to cater for when the population captures are not perfect, in that as well as there being some dependence there might be some capture error. The definition of perfect captures here is that firstly, any erroneous enumerations have been previously identified and resolved. Secondly, any resulting dependence is only attributable to list dependence; in other words, suitable post-strata have been chosen such that within strata capture probabilities are homogeneous. The approach for the estimation of population totals when there is imperfect data relies on latent class modelling. This lies within the general conceptual framework of latent variable models. Here it is believed that each individual's behaviour is conceived as being governed by their inherent (and therefore unobserved) traits.

In fact, the belief is that the observed systematic patterns in the population are better explained by some unobserved characteristics. By considering these unobservable characteristics into the modelling process, more often than not the inter-relationships become more clearer, and can therefore lead to better inference. The latent variable techniques - e.g. factor analysis, principal component analysis and discriminant analysis - effectively seek to reduce dimensionality so that by looking at the inter-relationships at the lower dimension patterns are more easy to detect and distinguish and the ability of the data analyst to see the structure of data is enhanced. However, although this basic idea is the same one underlying latent class analysis, the main difference is that these traits are discrete and distinct classes, and when further thought of as cells of a multi-dimensional contingency table nicely leads to the latent class model to be specified as a log-linear model.

In the previous chapters, the log-linear estimation models used assume that the cell counts had been removed of any erroneous enumerations. However, when there are erroneous enumerations, the modelling framework has to be amended, and this is accomplished through the inclusion of a latent variable that represents the enumeration status of each person. In brief, this chapter shows that the specification of the capture-recapture substantive problem as a log-linear model readily lends itself to easily cope with dependence and erroneous enumerations. Furthermore it will also be shown that if there is some additional association between the Census, Survey or Third List that is not fully accounted for by the latent variable then these residual direct effects can easily be incorporated into the log-linear model.

It has been previously demonstrated that the EM algorithm can be used to find estimates of the missingness, specifically when there is no measurement error. The EM algorithm is the general approach to finding maximum likelihood estimates in incomplete data problems, and has been the advocated method used during the thesis, though pages 237-242 of Bishop, Fienberg and Holland (1975) showed that closed form estimates exist for all of the log-linear models when there are perfect captures (i.e. there are no erroneous captures). The idea of using the EM algorithm means that it can be extended without much difficulty to the case when there is imperfect data from the three captures (and this imperfection could result either from a failure to correctly post-stratify or through the failure to remove erroneous enumerations). However, it must be noted that when there is imperfect data the complete likelihood includes an unobservable variable, and the EM algorithm has to be suitably modified. Nonetheless, the E and M steps are still performing the same functions of taking the expectation of the complete data likelihood conditional on the observed data augmented for some starting values of the missing data, and then computing new estimates of the missing data that maximize the likelihood. There is an added advantage of the EM algorithm in that the related result that there is a simple relationship between the complete, observed and missing data can be used. So here the complete data information is the sum of the observed data information and the missing data information (or put differently, the increase in variance due to missing data). This, in theory, allows the asymptotic variance-covariance matrix to be computed using the SEM algorithm, without the need of matrix inversions which as well as being complicated can lead to intractability.

In this chapter, an identifiable latent model that copes with both dependence and capture error is first presented, following on from the discussion at the end of Chapter 3. It will be shown that the latent model makes it possible to examine the validity of the no-three-way interaction assumption. The estimation of this model by way of the EM algorithm is also discussed. Section 6.5 presents the results of a feasibility study carried out that investigated the viability of fitting a latent model to some data with simulated capture error and dependence. Section 6.6 presents an alternative interpretation to the US Census data considered in Chapter 5 by fitting a latent class model. Sections 6.7

and 6.8 detail techniques available for providing precision estimates. In the estimation of the standard errors, Section 6.7 extends the SEM algorithm to the case when there is latentness. However as will be discussed, there are some problems with this extension of the SEM algorithm, due mostly to the identifiability of the latent model.

According to Goodman (1974) the best method of determining whether a model is identifiable or not relies on the Hessian matrix. The Hessian is the matrix of the second partial derivative of all free independent estimated parameters in the log-likelihood. It can be noted that the inverse of the Hessian matrix approximates the variance-covariance matrix of the parameter estimates. If the Hessian matrix has less than full rank, then the model is not identified. Notably it will be established here that albeit the models are indeed identified, there is not enough information to estimate some of the model parameters, and as such the SEM algorithm runs into difficulty when trying to estimate the asymptotic variance-covariance matrix. The bootstrap, on the other hand, does not suffer from these computational difficulties to the same degree, and results of the bootstrap-computed measures of precision for the Feasibility Study and US Census application data are duly presented in Section 6.8. Finally, the last part of the chapter briefly describes how population estimates are produced for non-sampled areas, since the second list is in fact a sample.

6.2 Using Latent Models to Cope with both Dependence and Capture Error

So far it has been shown that in triple system estimation the only assumption needed is that there is no second order interaction. This means that for the 'saturated model', all pairs of sources may exhibit dependence but the amount of dependence is assumed to have no bearing when conditioned on the third remaining source. This assumption can be thought of as analogous to the assumption of independence made in dual system estimation, in that it is untestable in isolation. In dual system estimation the only way of testing the assumption of independence relies on bringing in additional information, and similarly so in triple system estimation.

In the triple system models being considered during the Simulation Study in Chapter 4, there was rarely the need to question the assumption of no second order interaction, as the 'saturated' model with all three interaction effects overfitted the simulated data. However, one problem encountered in both the dual system and triple system estimators is that there is the need for the further assumption of error-free measurement of the population. This can be too heroic an assumption, especially when there are issues surrounding data collection, matching and biased response. As explained in Chapter 3, latent class modelling is being proposed to account for erroneous enumerations in the population measures.

It was also mentioned in Chapter 3 that the standard latent class model assumes that the population is composed of mutually exclusive latent classes such that within these classes the observed variables are unrelated; this is what was defined as local independence. This assumption is violated when there is reason to believe that, notwithstanding the relationships between the latent variable and observed variables, there are some relationships between the observed variables. This is particularly true for the type of latent class model under consideration when there is some additional correlation between the Census and Survey that will not be fully accounted for by the latent class model that pre-supposes local independence. Even so if the latent model is formulated as a log-linear model, it becomes possible to include the interaction between the Census and Survey

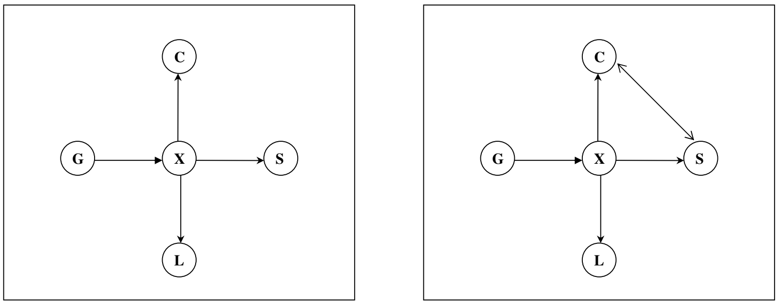
$$\log \mu_{ijkt} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)}. \tag{6.1}$$

This model, is however not identified since it has too many parameters. To circumvent these issues of identifiability the proposal (detailed earlier in Section 3.11 and Section 4.4) is to use additional covariates. So for example the case with just one covariate G is given by the following equation:

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)} + \lambda_{ig}^{(CG)} + \lambda_{jg}^{(SG)} + \lambda_{kg}^{(LG)} + \lambda_{gt}^{(GX)}. \tag{6.2}$$

If the covariate G is a dichotomous variable then it can be seen that (6.2) is exactly identified. Under the Goodman parameterization the local dependence model has the three-way interaction term CSX, which is clearly not identified, and it is not possible to write the log-linear model (6.2) in the Goodman parameterization without additional equality constraints. In fact the log-linear model (6.2) may actually be over-fitting the data, and a more parsimonious model, pictorially represented in Figures 6.1(a) and 6.1(b), might suffice for a carefully chosen covariate G. In effect the covariate G is only related to the latent variable, and this will prove to be very important when it comes to model identifiability.

Figure 6.1: Latent Class models with a covariate effect, G



(a) Local Independence

(b) Local Dependence

The interpretation of Figure 6.1(b) is that in order to fit the identifiable local dependence model the only requirement is that the CSX interaction term is zero. This preferred model can be clearly represented as the hierarchical log-linear model $\{CX, SX, LX, GX, CS\}$. Furthermore, this model is much simpler because it does not have the interaction terms between between the covariate G and the manifest variables. It is, however, a non-trivial matter to find such a grouping covariate, G. This is because G has to be chosen such that it is only related to the latent variable, X, but not the manifest variables, C, S and L.

Nevertheless, assuming such a covariate exists, then it is possible to fit the model to the contingency table of counts using the E- and M- steps. A further advantage of this additional covariate is that there are now enough degrees of freedom available to cope with the missing cell(s). So for a grouping covariate G with two levels, there are now 14 observed cells, and the latent local independence model to be fitted to the data is

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)},$$

and the corresponding local dependence latent model is

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)} + \lambda_{ij}^{(CS)}.$$

6.3 Fitting the Log-linear Latent Class Model to Triple System Data

It now remains to demonstrate if it is possible to fit a log-linear model to simulated data with the purpose of recovering

(a) the missing counts, and

(b) the latent classes.

The simulated data is obtained by generating expected cell counts for the 2^5 contingency table under local independence,

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)},$$

and local dependence,

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{gt}^{(GX)} + \lambda_{ij}^{(CS)}.$$

The expected cell counts are then collapsed over the latent variable, X, to produce a 2^4 contingency table. Finally, the cells corresponding to being missed on all three lists, $(0, 0, 0, g)$, are removed. For a correctly specified latent model, the belief is that it is possible to find both the missing and latent information given the $2^4 - 2$ observed cells. It must be mentioned that in the following Feasibility Study the data have been generated in such a way that they fit the latent model perfectly.

In the implementation of the EM algorithm described in Chapter 3, the original idea to cope with both dependence and erroneous enumerations was to conduct a two-stage process. So, given an observed contingency table, the first stage finds estimates of the missing cell counts, (n_{000g}). After this, the second stage finds estimates of the latent classes. However, this way of proceeding is proved to be incorrect. The reason why this leads to incorrect inference will be laid out shortly, but it is mainly due to how the E and M steps are being performed. To that end, in the next section the EM algorithm is modified to take account of both latentness and missingness. This is accomplished through the program **EM.latent** (see Appendix B.3), written in SPLUS/R, which fits different log-linear models to the observed data and then iteratively finds the maximum likelihood estimates.

In order to explore why the two-stage procedure fails, a Feasibility Study was carried out by slightly modifying the Simulation Study undertaken in Chapter 4 by placing erroneous enumerations in some specific cells. The aim is to then see if **EM.latent** was effective in correctly identifying these erroneous enumerations. As a simple example, suppose that the Census, Survey and Third List are mutually independent and all manage to achieve a 70% coverage rate of a population of size 1000. Also suppose there are 50 erroneous enumerations in total confined to the three cells (0,0,1), (1,0,0) and (1,0,1). Then the expected counts in the contingency table cells are $(n_{000}, 76, 63, 147, 88, 159, 147, 343)$. The data has been generated under independence, so it is expected that the local independence model will suffice, which is identifiable and therefore there is no need for a grouping covariate G.

When the independence model is fitted to these data then the estimate of the missing cell count, $\hat{n}_{000}$ equals 34. This estimate is not entirely correct since some of the observed counts are erroneous and this duly has an influence on the missing cell (in that the missing cell is overestimated). Consequently, it remains to see if **EM.latent** can correctly split this observed 2x2x2 table contingency table of counts into Erroneous and Real enumerations, the results of which appear in Table 6.1. It can be seen from Table 6.1 that the observed erroneous and real enumerations are different from the fitted values.

Table 6.1: Results of the fitted latent class model to the simulated data

Cell	Observed Real	Fitted Real	Observed Erroneous	Fitted Erroneous
n_{000}	27	29.17	7	4.83
n_{100}	63	76.00	12	0.00
n_{010}	63	60.81	0	2.19
n_{110}	147	147.00	0	0.00
n_{001}	63	65.66	25	22.34
n_{101}	147	159.00	13	0.00
n_{011}	147	136.90	0	10.10
n_{111}	343	343.00	0	0.00

An explanation of why they are so discrepant will be given shortly. But first consider another example. This time assume that a simulated population is generated with real and erroneous enumerations, such that 10% of the people are known to be erroneous. Also suppose that given that a person is real, then the probability that they will be counted in the Census, Survey or Third List is respectively 0.80, 0.90 and 0.60. On the other hand the probability that an erroneous person is counted in the Census is 0.15, and the corresponding probabilities in the Survey and Third List are 0.05 and 0.20. After marginalising over the latent variable the 2x2x2 (including the (0,0,0)-cell) counts expected are found to be $(\mathbf{71.80}, 40.20, 68.20, 259.80, 26.95, 46.05, 98.05, 388.95)$. Table 6.2, below, gives the results, and shows that the fitted values under the latent model correspond to the observed values.

Table 6.2: Results of the fitted latent class model to the simulated data

Cell	Observed Real	Fitted Real	Observed Erroneous	Fitted Erroneous
n_{000}	7.20	7.1999	64.60	64.6001
n_{100}	28.80	28.7998	11.40	11.4002
n_{010}	64.80	64.7990	3.40	3.4001
n_{110}	259.20	259.1998	0.60	0.6002
n_{001}	10.80	10.7999	16.15	16.1501
n_{101}	43.20	43.1999	2.85	2.8501
n_{011}	97.20	97.1997	0.85	0.8503
n_{111}	388.80	388.8000	0.15	0.1500

The reason why the latent model is able to find the correct number of real and erroneous enumerations under the second scenario but not in the first, relies on understanding that under latent class analysis the fundamental assumption is that after conditioning on the latent variable the relationship between the Census, Survey and Third List is the same in each of the latent classes. In other words, the relationships observed among the manifest variables (Census, Survey and Third List) is found to be the same, within the categories of the latent variable (correct enumeration status). This is the definition of local independence of the manifest variables on the latent variable, which underlies latent class modelling. For the data given in the first scenario since the erroneous enumerations can only occur in certain cells, it is not possible to accurately identify the real and erroneous enumerations by simply multiplying the conditional response probabilities and so the proposed latent class model does not fit the data, whereas this is possible in the second scenario. Incidentally, the n_{000} cell count in the second scenario of $\mathbf{71.80}$ cannot be reproduced by any log-linear model, *apart from* the fully saturated model for the complete data

$$\log \mu_{ijk} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_{ij}^{(CS)} + \lambda_{ik}^{(CL)} + \lambda_{jk}^{(SL)} + \lambda_{ijk}^{(CSL)}. \tag{6.3}$$

This means that the assumption of no three-way interaction that is pivotal to being able to fit the latent class model in the presence of the missing n_{000} cell does not hold here. In other words, it is not possible to use only the structure in the seven cell counts - $\{n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110}, n_{111}\}$ - to estimate the n_{000} cell count. This is because under model (6.3) the seven 'observable' cells are not sufficient for the estimation of the missing cell. Therefore, the relationship between the 'observable' cells is not sufficient to find an unbiased estimate of the population size. Put simply, the (0,0,0)-cell provides some additional information, not contained in the other cells.

The proof of this result appears in Chapter 5 of Salgueiro (2002), but relies on the premise that the observed contingency table is not fully represented by the observed variables. Basically this implies that there exists a latent variable such that the observed variables are conditionally independent given that variable, and this latent model is the most sensible. (Salgueiro, 2002, page 187-189) showed that marginalising over this latent variable induces the saturated model, and no other model. Thus, the most adequate way of representing the associations and interactions between the observed variables is only through the model that includes **all** the terms, i.e. model (6.3).

This is the conundrum faced in capture-recapture methods; the over-riding assumption is that the information provided by the observed cells should provide more than an adequate insight into what is happening in the unobserved cell, but it is untestable, so when it does not hold, any solution arrived at could potentially be wrong. Additionally, since the no three-way interaction is deemed not to be the most parsimonious model, the expectation is that there exists a simpler, less complicated, model that explains much of the variation in the observed cells. Admittedly, this is not too unreasonable an assumption to make under the circumstances when there is no overenumeration. However the above has demonstrated that when there is, the no-three-way interaction assumption is invalidated. This means that a two-step process of using the observed data ($2^r - 1$ cells) to estimate the missing cell, and then fitting a latent class model to the 2^r cells with an estimate of the missing cell included does not work. The favoured approach will be to fit the latent class model directly to the observed cells, and then to iteratively find the missing cell estimate. The intent is that at convergence each observed cell divides into 'real' and 'erroneous' counts, and subsequently the cell corresponding to 'real' missed and 'erroneous' missed people can be derived.

6.4 An Identifiable Latent Log-linear Model

The preceding section has shown that there are just enough degrees of freedom available to fit the latent log-linear model

$$\log \mu_{ijkt} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)}.$$

However, in reality the (0,0,0)-cell is never observed and so the latent log-linear model becomes un-identified. During the course of the thesis, it was suggested that a two-stage process could be undertaken. Thus, starting with the seven observable cells, a suitable model was proposed and an estimate of the missing cell found. Next, a latent log-linear model was then fitted these eight-cells (the seven observed plus the estimated missing cell) so that it split the data into real and erroneous enumerations. Nonetheless, for data generated with a latent variable, which was subsequently marginalised over to obtain a 2x2x2 contingency table it was demonstrated in the previous section that no model apart from the one with the CSL interaction effect managed to reproduce the correct estimate of the (0,0,0)-cell; this model is unfortunately not identified. The proposed solution to this problem is to come up with a covariate G, whose relationship to the manifest variables is only through the latent variable. So supposing, such a grouping covariate could be found it was suggested that the EM algorithm can be implemented directly to the incomplete contingency table to find missing cell, even when the lists are imperfect.

As such, the EM algorithm starts by some initial values $\mu^{(0)}_{ijkgt}$ and then the M-step fits the log-linear model

$$\log \hat{\mu}_{ijkgt} = \hat{\lambda} + \hat{\lambda}^{(C)}_i + \hat{\lambda}^{(S)}_j + \hat{\lambda}^{(L)}_k + \hat{\lambda}^{(G)}_g + \hat{\lambda}^{(X)}_t + \hat{\lambda}^{(CX)}_{it} + \hat{\lambda}^{(SX)}_{jt} + \hat{\lambda}^{(LX)}_{kt} + \hat{\lambda}^{(GX)}_{gt}. \qquad \text{(M)}$$

Now, given the data n_{ijkg}, with n_{000g} unobserved, the E-step consists of two sub-steps. Firstly an estimate of the missing cells is obtained for each group g,

$$\hat{n}_{000g} = \sum_t \hat{\mu}_{000gt}, \qquad \text{(E1)}$$

i.e. resulting in a 'full' observed contingency table. Then, secondly the latent cells are estimated by

$$\hat{n}_{ijktg} = \frac{n_{ijkg}}{\hat{\mu}_{ijkg}} \hat{\mu}_{ijkgt}. \qquad \text{(E2).}$$

This process of computing the expectation of the complete data likelihood conditional on the observed data, when repeated will converge to a solution that maximizes the expected likelihood. It was found that the convergence of this algorithm was quite similar to the case when the n_{000g} cells are assumed to be observed.

6.5 Feasibility Study into the Use of Latent Class Models in Population Size Estimation from Imperfect Data

As a simple demonstration of how this log-linear latent model works, data were generated, initially under local independence and then under local dependence. Note that the local independence model using the 2x2x2 contingency table was previously found to be unidentified when the cell count, n_{000}, was missing. The EM algorithm described in Section 6.4 can be used to fit a latent log-linear model in the presence of missingness, with the help of an additional covariate that is not directly correlated with the manifest variables. Thus it is surmised that the observed variables do not influence each other directly, but their inter-relationships are entirely derived from the correlation of the joint variable CS and the variable L with the latent variable X.

So here first the simulation finds the latent probabilities given the conditional probabilities $\pi_{it}^{C|X}$, $\pi_{jt}^{S|X}$ and $\pi_{kt}^{L|X}$ for the case with local independence, and $\pi_{ijt}^{CS|X}$ and $\pi_{kt}^{L|X}$ when there is local dependence. Then, the grouping variable probabilities π_{g}^{G} and $\pi_{gt}^{G|X}$ are supplied. Subsequently, the probabilities that a randomly selected case will be located in cell (i, j, k, g, t) is given by,
under local independence,

$$\pi_{ijkgt}^{CSLX} = \pi_{g}^{G}\pi_{gt}^{G|X}\pi_{it}^{C|X}\pi_{jt}^{S|X}\pi_{kt}^{L|X},$$

and under local dependence,

$$\pi_{ijkgt}^{CSLX} = \pi_{g}^{G}\pi_{gt}^{G|X}\pi_{ijt}^{CS|X}\pi_{kt}^{L|X}.$$

The first scenario considered is the simple case of local independence. As usual, a simulated population of 1000 is considered, and in this population assume that there is a latent variable that denotes whether or not a person is real or erroneous. Again, given that a person is real the probability of being counted in the Census, Survey and Third List are 0.80, 0.90 and 0.60 respectively. Similarly, given that a person is erroneous then the probability counted in the Census, Survey and Third List are 0.15, 0.05 and 0.20. The only difference with the previous simulations is that the population has been subdivided into two groups of which a third are in Group 1 and the remaining two-thirds are in Group 2. To further demonstrate the flexibility of the EM algorithm, it is assumed that the proportion of erroneous enumerations in the two groups are different; with there being 10% in Group 1, and 20% in Group 2 (so the erroneous enumerations make up 16.7% of the population). Now when the latent variable is marginalised over, the 2x2x2x2 contingency table results with expected cell counts (**23.91**, 13.39, 22.71, 86.51, 8.97, 15.33, 32.65, 129.52, **90.45**, 32.28, 42.95, 154.48, 27.95, 29.41, 58.76, 230.72). The missing cells (in bold) are then removed from the contingency table. Evidently since the n_{000g} cells cannot be observed, the aim is two-fold, to see if it is possible reproduce the missing cells, and also split the data into the latent contingency table. From the Table 6.3, it can be seen that the EM algorithm is effective in correctly identifying the latent classes as well as managing to find the missing.

Table 6.3: Results of the identifiable latent class model to the simulated data - local independence

Cell	Observed Real	Fitted Real	Observed Erroneous	Fitted Erroneous
$\hat{n}_{000g_1}$	**2.40**	**2.3975**	**21.51**	**21.5084**
n_{100g_1}	9.59	9.5902	3.80	3.7964
n_{010g_1}	21.58	21.5783	1.13	1.1323
n_{110g_1}	86.31	86.3135	0.20	0.1999
n_{001g_1}	3.60	3.5963	5.38	5.3781
n_{101g_1}	14.39	14.3854	0.95	0.9493
n_{011g_1}	32.37	32.3675	0.28	0.2831
n_{111g_1}	129.47	129.4704	0.05	0.0500
$\hat{n}_{000g_2}$	**4.27**	**4.2687**	**86.18**	**86.1613**
n_{100g_2}	17.08	17.0748	15.21	15.2080
n_{010g_2}	38.42	38.4190	4.54	4.5358
n_{110g_2}	153.68	153.6766	0.80	0.8006
n_{001g_2}	6.40	6.4030	21.54	21.5443
n_{101g_2}	25.61	25.6120	3.80	3.8027
n_{011g_2}	57.63	57.6285	1.13	1.1342
n_{111g_2}	230.52	230.5151	0.20	0.2002

In the second scenario where there is local dependence, the assumption is that the Census and Survey are independent of the Third List (so although the effect of the Census varies across different levels of the Survey, the effect of the Third List remains unchanged). Consequently, the Census and Survey can be thought of as a single (joint) variable with four levels $\{00, 01, 10, 11\}$. Thus, given a person was erroneous the probability that they were found by both the Census and Survey, found by the Census and missed by the Survey, missed by the Census and found by the Survey was respectively 0.3, 0.2 and 0.1. Since the probabilities must sum to 1, it follows that the conditional probability that an erroneous person was missed by both the Census and Survey was 0.4. Similarly, given that a person was real the four probabilities were 0.3, 0.2, 0.4 and 0.1. The Third List conditional probabilities are kept the same as for the local independence case; i.e. 0.6 for real enumerations and 0.2 for erroneous enumerations. Also keeping the grouping subdivisions (i.e. $\frac{1}{3}$ in Group 1 and $\frac{2}{3}$ in Group 2) and the proportion of erroneous in the groups the same as above, the resulting 2x2x2x2 contingency table of expected counts is (**31.97**, 31.97, 47.95, 47.95, 21.31, 34.63, 65.27, 51.95, **45.36**, 58.70, 101.38, 88.04, 41.35, 74.70, 145.41, 112.06).

Specifically when there is local dependence, the only part of the EM algorithm that changes is the M step, with the model fitted at the new M step given by

$$\log \hat{\mu}_{ijkgt} = \hat{\lambda} + \hat{\lambda}_i^{(C)} + \hat{\lambda}_j^{(S)} + \hat{\lambda}_k^{(L)} + \hat{\lambda}_g^{(G)} + \hat{\lambda}_t^{(X)} + \hat{\lambda}_{ij}^{(CS)} + \hat{\lambda}_{it}^{(CX)} + \hat{\lambda}_{jt}^{(SX)} + \hat{\lambda}_{kt}^{(LX)} + \hat{\lambda}_{gt}^{(GX)} + \hat{\lambda}_{ijt}^{(CSX)}. \tag{M}$$

The observed and fitted results are given in Table 6.4 and show broad agreement, and most of the differences between them could be attributed to rounding and tolerance, which is to be expected given that the EM algorithm is fitting the correct model under the simulated data.

Table 6.4: Results of the identifiable local dependence latent model - simulated data with different CSX table odd ratios

Cell	Observed Real	Fitted Real	Observed Erroneous	Fitted Erroneous
$\hat{n}_{000g_1}$	**10.66**	**10.6219**	**21.31**	**21.2844**
n_{100g_1}	21.31	21.3052	10.66	10.6838
n_{010g_1}	42.64	42.6318	5.33	5.3382
n_{110g_1}	31.97	31.9578	15.98	15.9942
n_{001g_1}	15.98	15.9798	5.33	5.3322
n_{101g_1}	31.97	31.9608	2.66	2.6712
n_{011g_1}	63.94	63.9360	1.33	1.3374
n_{111g_1}	47.95	47.9412	4.00	4.0068
$\hat{n}_{000g_2}$	**24.01**	**24.0068**	**21.34**	**21.3265**
n_{100g_2}	48.02	48.0122	10.67	10.6838
n_{010g_2}	96.05	96.0531	5.34	5.3489
n_{110g_2}	72.04	72.0274	16.01	16.0257
n_{001g_2}	36.02	36.0113	5.34	5.3427
n_{101g_2}	72.04	72.0183	2.67	2.6766
n_{011g_2}	144.07	144.0660	1.33	1.3400
n_{111g_2}	108.05	108.0412	4.00	4.0148

Haberman (1979) intimated that the latent class model was very sensitive to the correct model specification, and this is proved when a different maximization step was used, this time removing the CSX interaction. It was hoped that the $\{CX, SX, LX, CS\}$ model was a parsimonious representation of the $\{LX, CSX\}$ model. Hence, the new M step for this simpler model is

$$\log \hat{\mu}_{ijkgt} = \hat{\lambda} + \hat{\lambda}_i^{(C)} + \hat{\lambda}_j^{(S)} + \hat{\lambda}_k^{(L)} + \hat{\lambda}_g^{(G)} + \hat{\lambda}_t^{(X)} + \hat{\lambda}_{ij}^{(CS)} + \hat{\lambda}_{it}^{(CX)} + \hat{\lambda}_{jt}^{(SX)} + \hat{\lambda}_{kt}^{(LX)} + \hat{\lambda}_{gt}^{(GX)}. \tag{M}$$

The $\{CX, SX, LX, CS\}$ model was however found to be a fairly poor fit to the data. An explanation for why this could be is found on examining the odds ratios of the CSX marginal probabilities, as shown in Table 6.5. The two-way interaction between the Census and Survey, CS, varies across levels of the latent variable, X. For $X = 1$, i.e. erroneous enumerations, the odds ratio is 6, while the corresponding odds ratio when $X = 2$ is $\frac{3}{8}$. It follows that since the odds ratios are different, it is to be expected that the CSX interaction has a significant effect on the model fit.

Therefore, another simulation was carried out, this time the CSX marginal probabilities were chosen such that odds ratios for $X = 1$ and $X = 2$ were the same (see Table 6.6). The resulting n_{ijkg} table of counts was (**26.62**, 42.62, 53.28, 37.30, 19.98, 37.30, 66.60, 49.28, **40.02**, 69.37, 106.72, 77.37,40.02, 77.37, 146.74, 109.39) with the missing cell counts in bold. Again the objective here is to obtain estimates of these missing cells with an adjustment for the unobserved latent variable. However, this time there were 20% erroneous enumerations in Group 1 and 10% in Group 2 (so overall there are 13.3%

Table 6.5: Census, Survey, Latent (CSX) Marginal Table I

		Erroneous , X=1			Real, X=2	
		Survey			Survey	
		Yes	No		Yes	No
Census	Yes	0.3	0.2	Yes	0.1	0.4
	No	0.1	0.4	No	0.2	0.3
		Odds Ratio=6			Odds Ratio=$\frac{3}{8}$	

Table 6.6: Census, Survey, Latent (CSX) Marginal Table II

		Erroneous , X=1			Real, X=2	
		Survey			Survey	
		Yes	No		Yes	No
Census	Yes	0.1	0.4	Yes	0.3	0.2
	No	0.2	0.3	No	0.4	0.1
		Odds Ratio=6			Odds Ratio=6	

erroneous enumerations in the population), but every other simulation parameter was kept the same as before.

The EM algorithm to fit the model to the data simulated with the same marginal CSX probabilities now uses the M-step maximizing the expected likelihood under the model $\{CX, SX, LX, GX, CS\}$. Table 6.7 shows the fitted and observed counts - and as anticipated there is broad agreement. Moreover in the analysis, it was found that fitting either the $\{CX, SX, LX, GX, CS\}$ or $\{GX, LX, CSX\}$ model produced the same results.

Furthermore, the good thing about bringing the covariate into the latent class model is that the choice can be made between the $\{CX, SX, LX, GX, CS\}$ and $\{GX, LX, CSX\}$ models - both models are identifiable. If the two models give different results then, it is better to choose the more complicated model (but it must be noted that this model can take a little while longer to converge). The simulations have shown that the latent EM algorithm works only when the correct M-step is used for the data at hand, but how can it be possible, given a 2x2x2x2 contingency table with missing n_{000g} cells, to fit the correct model? This is because when there is missingness it is not possible to look at the odds ratios in order to decide between fitting a model with local independence and local dependence. The simplest and most effective way to detect local dependence, as suggested by McCutcheon (1987) and Hagenaars (1993), is to look at the goodness of fit statistics

$$X^2 = \sum_i \sum_j \sum_k \sum_g \frac{(n_{ijkg} - \hat{\mu}_{ijkg})^2}{\hat{\mu}_{ijkg}} \quad \text{or} \quad G^2 = 2 \sum_i \sum_j \sum_k \sum_g n_{ijkg} \log \left(\frac{n_{ijkg}}{\hat{\mu}_{ijkg}} \right),$$

for $(ijkg) \neq (000g)$.

If there is no difference between the model fitted under the local independence or local dependence assumption, then as well as having the same parameter estimates, the model

Table 6.7: Results of the identifiable local dependence latent model - simulated data with the same CSX table odd ratios

Cell	Observed Real	Fitted Real	Observed Erroneous	Fitted Erroneous
$\hat{n}_{000g_1}$	**10.66**	**10.6560**	**15.99**	**15.9840**
n_{100g_1}	21.31	21.3119	21.31	21.3441
n_{010g_1}	42.62	42.6239	10.66	10.6561
n_{110g_1}	31.97	31.9679	5.33	5.3281
n_{001g_1}	15.98	15.9840	3.40	3.3960
n_{101g_1}	31.97	31.9679	5.33	5.3361
n_{011g_1}	63.94	63.9359	2.66	2.6681
n_{111g_1}	47.95	47.9520	1.33	1.3340
$\hat{n}_{000g_2}$	**24.01**	**24.0119**	**16.01**	**16.0080**
n_{100g_2}	48.02	48.0240	21.34	21.3441
n_{010g_2}	96.05	96.0478	10.67	10.6721
n_{110g_2}	72.04	72.0359	5.34	5.3361
n_{001g_2}	36.02	36.0180	4.00	4.0020
n_{101g_2}	72.04	72.0359	5.34	5.3361
n_{011g_2}	144.07	144.0720	2.67	2.6681
n_{111g_2}	108.05	108.0540	1.33	1.3340

Table 6.8: Results of the different class latent models fitted to the simulated data

		Model 1	Model 2	Model 3
	Observed	Estimated	Estimated	Estimated
n_{0001}	NA	**166.12**	**26.62**	**25.63**
n_{1001}	42.62	41.36	42.62	42.56
n_{0101}	53.28	55.28	53.28	53.34
n_{1101}	37.30	37.20	37.30	37.36
n_{0011}	19.98	27.60	19.98	19.98
n_{1011}	37.30	27.46	37.30	37.41
n_{0111}	66.60	58.88	66.60	66.51
n_{1111}	49.28	58.58	49.28	49.19
n_{0002}	NA	**249.63**	**40.01**	**39.28**
n_{1002}	69.37	70.82	69.37	69.45
n_{0102}	106.72	107.00	106.72	106.65
n_{1102}	77.37	77.97	77.37	77.23
n_{0012}	40.02	59.85	40.02	40.02
n_{1012}	77.37	59.55	77.37	77.23
n_{0112}	146.74	127.71	146.74	146.84
n_{1112}	109.39	127.06	109.39	109.49
X^2		25.465	0	0.002
G^2		19.841	0	0.007
df		4	3	2
p-value		0.0006	1	1

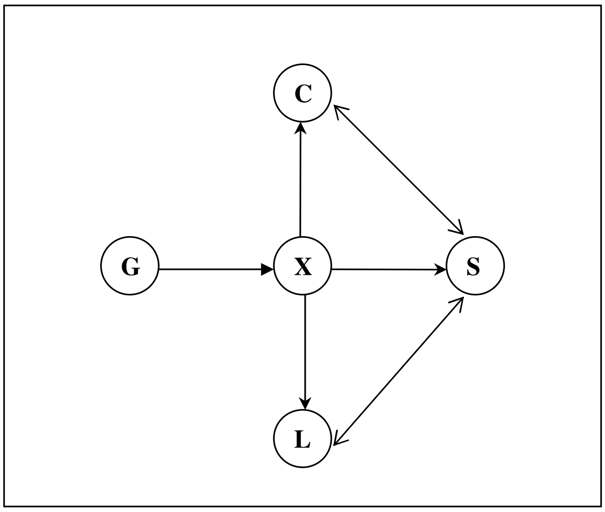
fit statistics will be practically the same. Table 6.8 illustrates how well the different models fit the simulated data. Model 1 is the local independence model $\{CX, LX, SX, GX\}$, while Models 2 and 3 are the local dependence models $\{CX, LX, SX, GX, CS\}$ and $\{GX, LX, CSX\}$ respectively. The goodness of fit statistics have been calculated excluding the n_{000g} cells. Nonetheless, it can be seen that the n_{000g}-cell estimates vary widely under local independence and local dependence, with there being a massive over-estimation of the these cells under local independence. In this scenario, it is known that the data have been simulated under local dependence so the model assuming local independence is rejected. At least a look at the goodness of fit statistics X^2 and G^2 confirms this, since the local independence model provides a poor fit to the simulated data.

6.6 Fitting Latent Class Models to the US 1990 Census Dress Rehearsal Data - a different interpretation?

The discussion that came out of the analysis of the US Census data in the previous chapter was about the seeming difference in the estimates of the missing counts between what was deemed to be the best-fitting model, $\{CS, SL\}$ and the 'saturated model', $\{CS, CL, SL\}$. In the analysis in Chapter 5, there was found to be a very large difference between these two models, and for some post-strata the estimate under the $\{CS, CL, SL\}$ model was almost three times the size of that under model $\{CS, SL\}$. Now for capture-recapture log-linear modelling, it can be remembered that, the over-riding assumption is that the most complicated model that can be fitted to the data is the homogeneous association model, meaning that the conditional odds ratios between any two variables are identical for each category of the third variable. Further, it is expected that there is a less complicated model that fits the data equally well. Under these conditions, the 'saturated' and best-fitting models are anticipated to yield the same estimates of the missing counts. The difference was of concern, although using the SEM-calculated variances, the 95% confidence intervals showed that for most of the post-strata either estimates from the two competing models were plausible.

It was also suggested in Chapter 5 that the reason for the ambiguity in the conclusions - i.e. two different models seem to fit the model well - could be due to a latent variable, which was exhibited by the failure of the no-three-way interaction assumption. Also due to the way in which the data were collected, it would make substantive sense to include interaction effects between the Census and Survey and the Survey and Administrative List. In other words, the latent variable does not fully account for all the dependence between the Census, Survey and List. In fact there is still some residual dependence, and as such the local dependence model is required. Obviously, this model is not identifiable, since the model is over-parameterised. The suggested solution of bringing a grouping covariate into the frame, such that the effect of each the manifest variables is mediated through the latent variable, pictorially represented in Figure 6.2, brings about identifiability.

Figure 6.2: Path diagram of the local dependence model with two direct effects.



This latent class model is given by

$$\log \mu_{ijktg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)} + \lambda_{gt}^{(GX)}.$$

and is identified, since G has four levels - O2, R2, O3 and R3 - there are 28 observations and 12 estimable parameters in the proposed model. It will also be remembered that the 'best fitting' model in Section 5.2

$$\log \mu_{ijkg} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_g^{(G)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)} + \lambda_{ig}^{(CG)} + \lambda_{jg}^{(SG)} + \lambda_{kg}^{(LG)} + \lambda_{ijg}^{(CSG)} + \lambda_{jkg}^{(SLG)}$$

has the three variable interaction terms, λ_{ijg}^{CSG} and λ_{jkg}^{SLG}.
Even so it is possible to include these addtional interaction terms to the model leading to the latent model

$$\begin{aligned} \log \mu_{ijktg} = \quad & \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_t^{(X)} + \lambda_g^{(G)} + \lambda_{it}^{(CX)} + \lambda_{jt}^{(SX)} + \lambda_{kt}^{(LX)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)} \\ & + \lambda_{gt}^{(GX)} + \lambda_{ig}^{(CG)} + \lambda_{jg}^{(SG)} + \lambda_{ijg}^{(CSG)} + \lambda_{jkg}^{(SLG)} \end{aligned}$$

which is still identifiable.

A subsequent investigation was carried out to determine if fitting latent class models to the data could possibly illuminate this and try to validate the results presented in Chapter 5. Using the US Census Dress Rehearsal results presented in a 2x2x2x4 contingency table (see Table 5.6), with four missing cells (n_{000O_2}, n_{000O_3}, n_{000R_2} and n_{000R_3}), different latent models were fitted to the data and the results were found to be very interesting. The results from the analysis are found to yield more appealing interpretations than those of the log-linear modelling in Chapter 5. Table 6.9 gives the estimates of the two latent classes under different model specifications. The $\{LX, GX, CSG, SLG\}$ model produced results not too dissimilar to the $\{CS, SL, CX, SX, LX, GX\}$ model. The first thing of note is that the estimate of the missing cell in the second latent class is always zero, and this is so whichever way the model is specified, be it under local independence or the various specifications of local dependence.

Table 6.9: Latent Class Analysis of the US Census Data

	n_{000}	n_{100}	n_{010}	n_{110}	n_{001}	n_{101}	n_{011}	n_{111}
Local Independence								
Latent Class 1								
O2	**153.90**	31.00	5.86	1.22	59.00	19.00	6.31	1.46
R2	**159.05**	41.00	23.04	5.08	43.00	12.00	3.05	0.83
O3	**153.59**	62.00	6.81	2.68	35.00	13.00	2.80	1.32
R3	**129.07**	32.00	14.52	3.77	43.00	7.00	2.83	0.75
Latent Class 2								
O2	**0.00**	0.00	2.14	11.78	0.00	0.00	12.69	77.54
R2	**0.00**	0.00	10.96	63.92	0.00	0.00	7.95	57.17
O3	**0.00**	0.00	3.19	33.32	0.00	0.00	7.20	89.67
R3	**0.00**	0.00	9.48	65.23	0.00	0.00	10.17	71.25
Local Dependence (with CS interaction)								
Latent Class 1								
O2	**153.55**	31.00	5.84	0.00	59.00	19.00	6.57	0.00
R2	**153.84**	41.00	22.58	0.00	43.00	12.00	3.06	0.00
O3	**151.28**	62.00	6.74	0.00	35.00	13.00	2.88	0.00
R3	**125.41**	32.00	14.27	0.00	43.00	7.00	2.89	0.00
Latent Class 2								
O2	**0.00**	0.00	2.16	13.00	0.00	0.00	12.43	79.00
R2	**0.00**	0.00	11.42	69.00	0.00	0.00	7.94	58.00
O3	**0.00**	0.00	3.26	36.00	0.00	0.00	7.12	91.00
R3	**0.00**	0.00	9.73	69.00	0.00	0.00	10.11	72.00
Local Dependence (with SL interaction)								
Latent Class 1								
O2	**153.68**	31.00	7.22	1.63	59.00	19.00	14.52	3.76
R2	**161.05**	41.00	29.82	6.84	43.00	12.00	7.85	2.14
O3	**149.20**	62.00	8.69	3.34	35.00	13.00	6.98	3.13
R3	**132.68**	32.00	20.11	5.10	43.00	7.00	8.36	1.95
Latent Class 2								
O2	**0.00**	0.00	0.78	11.37	0.00	0.00	4.48	75.24
R2	**0.00**	0.00	4.18	62.16	0.00	0.00	3.15	55.86
O3	**0.00**	0.00	1.31	32.66	0.00	0.00	3.02	87.87
R3	**0.00**	0.00	3.89	63.90	0.00	0.00	4.64	70.05
Local Dependence (with CS and SL interactions)								
Latent Class 1								
O2	**153.34**	31.00	6.56	0.00	59.00	19.00	10.84	0.00
R2	**155.34**	41.00	26.14	0.00	43.00	12.00	5.42	0.00
O3	**149.35**	62.00	7.70	0.00	35.00	13.00	4.94	0.00
R3	**127.26**	32.00	17.06	0.00	43.00	7.00	5.43	0.00
Latent Class 2								
O2	**0.00**	0.00	1.44	13.00	0.00	0.00	8.16	79.00
R2	**0.00**	0.00	7.86	69.00	0.00	0.00	5.58	58.00
O3	**0.00**	0.00	2.30	36.00	0.00	0.00	5.06	91.00
R3	**0.00**	0.00	6.94	69.00	0.00	0.00	7.57	72.00

During the introduction of latent class analysis mention was made of the fact that the latent variable is brought in to account for unobserved heterogeneity. However, there could be two different interpretations of the latent classes (precisely, enumeration error and enumeration difficulty), which depending on the interpretation chosen will lead to a different population estimate; and as a result it was expressed that there could be two, conflicting, population estimates. The first interpretation is the one that the majority of thesis has focused on, while the second interpretation is based on catchability. Individuals are presumed to cluster into latent classes such that individuals within the same class have the same catchability. The basis of post-stratification is to ensure that subgroups of the population are chosen so that within each subgroup the individuals roughly exhibit the same capture behaviour, and the failure to correctly post-stratify leads to heterogeneity bias in the population estimates. It was stated at the beginning of the thesis that provided the population has been suitably post-stratified then fitting a latent model to the cross-classified table of counts leads to classes that represent enumeration error. For the US Census Dress Rehearsal data, however, the examination of the latent classes suggests that the unobserved heterogeneity could be attributable to enumeration difficulty and not enumeration error.

Moreover, the estimates of the missing (roughly 155 persons in O2, 155 in R2, 150 in O3 and 125 in R3) when compared to the results in Chapter 5 are somewhat closer to the best-fitting model (the one with the CS, SL interactions) whose corresponding estimates of missing were 96 people in O2, 147 in R2, 167 in O3 and 197 in R3, than the 'saturated' model. In Section 6.3 it was shown that when the untestable no-three-way interaction assumption cannot be justified - mainly due to the data being marginalised over a latent variable - there may be issues surrounding the correct estimation of the population size.

It emerges that every person who appears in the n_{001}, n_{100} and n_{101} cells, as well as those estimated to be missing (n_{000}), is placed in the first latent class, for all the models. On the other hand, every person who is counted in both the Census and Survey, i.e. n_{110} and n_{111} cells, is placed in the second latent class. It is also apparent the remaining cells, n_{010}, n_{011}, represent those people who were counted and it appears that for these cells the latent model distributes the observed people to the two classes by some mechanism. The interpretation of what the two latent classes actually represent is difficult, but a crude one that may be offered is that the latent classes denote whether or not a person was found on the Survey. In essence the observed contingency table data from the US 1990 Census Dress Rehearsal in St Louis shows a mixture of two latent subgroups - one group of people can be described as being easy to count by the Survey, and the other subgroup are hard to count by the Survey. Nevertheless, what this basically shows is that the data as they appears in Table 5.6 suffers from a failure of the post-stratification scheme, and as such there is some residual heterogeneity not fully accounted for by the age, race, tenure post-strata.

6.7 Problems with the Supplemented EM Algorithm for Standard Error Computation in Latent Class Analysis

Since the EM algorithm has been applied to compute the maximum likelihood estimates of the latent model parameters, it was anticipated that the asymptotic covariances and variances of the model parameters could be obtained using the SEM algorithm without a great deal of effort. However, the implementation of the SEM for latent class models did encounter some problems which will be discussed in this section. The idea of the SEM is a simple one and for the case when there is a single parameter to be maximized, the general formulation of the SEM algorithm, as detailed in Meng and Rubin (1991) is comprised of these steps:

1. Obtain an initial estimate $\theta^{(0)}$ that satisfies the log-linear model.

2. Run the EM algorithm to convergence to find the MLE, θ^*.

3. Define $\theta^{(t)}$ to be the EM estimate of θ at the tth iteration.

4. The rate of convergence, r, is given by
$$r = \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*}$$

5. Run the EM and calculate r until convergence is reached such that
$$r = \frac{\theta^{(t+1)} - \theta^{(t)}}{\theta^{(t)} - \theta^{(t-1)}}.$$

6. The observed variance can now be obtained from the known complete data variance, V_{com}, and the rate of convergence, r to be
$$V_{obs} = V_{com} + \Delta V = V_{com} + \frac{r}{1-r} V_{com}.$$

So similarly, for the case when the EM algorithm is maximizing over a multiple parameter space (a case in point is when there is unobserved latentness) then the above SEM algorithm can be suitably amended to

- find the correct matrix version of r, (r_{ij}) referred to as DM,
- find the correct matrix version of ΔV and
- find the correct matrix version of $V_{obs} = V_{com} + \Delta V$.

As usual, the SEM algorithm works by first obtaining the (multi-parameter) ML estimate $\boldsymbol{\theta^*}$ given some initial estimates $\boldsymbol{\theta^{(0)}}$. Define r_{ij} to be the $(i,j)^{th}$ element of the DM matrix.

1. At the t^{th} iteration let the EM estimate be $\theta^{(t)}$.

2. Define $\theta^{(t)}(i) = \left(\theta_1^*, \theta_2^*, \ldots, \theta_{i-1}^*, \theta_i^{(t)}, \theta_{i+1}^*, \ldots, \theta_d^*\right)$, which is effectively all the components fixed at their MLEs, except the i^{th} component, $\theta_i^{(t)}$.

3. Now run a single iterate of the EM algorithm, and find the $(i,j)^{th}$ element of the DM

$$r_{ij} = \frac{\theta_j^{(t+1)}(i) - \theta_j^*}{\theta_i^{(t)} - \theta_i^*} \qquad \text{for} \quad j = 1, \ldots, d.$$

Alternatively since DM is the Jacobian, r_{ij} is the $(i,j)^{th}$ term of the Jacobian

$$\begin{aligned} r_{ij} &= \frac{\partial M_j(\boldsymbol{\theta}^*)}{\partial \theta_i} \\ &= \lim_{\theta_t \to \theta^*} \frac{M_j\left(\theta_1^*, \theta_2^*, \ldots, \theta_{i-1}^*, \theta_i^{(t)}, \theta_{i+1}^*, \ldots, \theta_d^*\right) - M_j(\theta^*)}{\theta_i - \theta_i^*} \\ &= \lim_{t \to \infty} \frac{M_j\left(\theta^t(i)\right) - \theta_j^*}{\theta_i - \theta_i^*}. \end{aligned}$$

The procedure of running the EM steps and calculating the components of the DM is iteratively run until all of the r_{ij} are stable.

4. The complete data variance, V_{com} is simply the covariance-variance matrix of the parameters when the missing cells are directly substituted by the converged EM algorithm parameter estimates.

5. With DM and V_{com} evaluated, the observed covariance-variance can be computed as $V_{obs} = V_{com} + \Delta V = V_{com} + V_{com} DM \left(1 - DM\right)^{-1}$.

6. The asymptotic variances of the parameters are given by the diagonal terms of the matrix, V_{obs}, and the confidence intervals can now be derived as $\exp\left(\boldsymbol{\beta} \pm 1.96\boldsymbol{\sigma}\right)$, where $\boldsymbol{\beta}$ is the multi-parameter MLE, and $\boldsymbol{\sigma}$ is vector of the square root of the asymptotic variances.

In Chapter 5, it was shown that the SEM algorithm can be used to find the asymptotic covariance matrix of the model parameters obtained by the EM algorithm. Key to the derivation of the asymptotic covariance matrix is the DM matrix - even though the EM mapping $\theta^{(t+1)} \to M\left(\theta^{(t)}\right)$ does not have an explicit form, the derivative of the mapping DM can, in theory, be estimated through numerical methods. For the SEM implemented in Chapter 5, the DM is fairly straightforward to compute. However, for the case when there is a latent variable, the derivation of the DM can be difficult, possibly due to the complexity introduced by the latent information. Bearing in mind that the computation of the DM matrix involves essentially obtaining the Jacobian through numerical differentiation because of the inability to directly evaluate it, when the numerical differentiation is not close to the actual Jacobian there are inaccuracies in the SEM-derived covariances. This is often shown by the asymmetry in the resulting variance-covariance matrix at convergence. Another problem encountered in using the SEM when there is a latent variable is what is referred to as *weak identifiability* (see Garrett and Zeger (2000), Formann (2003)). During the SEM implementation, an indication of weak identifiability in a model was found to be exhibited by the V_{com} terms corresponding to the latent parameters being very large[1].

The best way to demonstrate weak identifiability is to consider an example where the observed contingency table is given by (170, 15, 0, 0, 6, 0, 0, 0, 4, 17, 0, 83, 1, 4, 0, 128). Clearly, it can be seen that the majority of people appear in the n_{000g1} and n_{111g2} cells. Without removing the n_{000g1} and n_{000g2} cells from the analysis, and fitting a local independence two-class latent model to the data the parameter estimates are shown in the second column of Table 6.10. As a consequence the complete data covariance matrix, V_{com} is easily computed by substituting the parameter estimates at convergence of the EM algorithm, and the diagonal elements give the variance of the parameters under complete information. The standard errors are the square roots of the variances, and are shown in the third column of Table 6.10. From here it can be noticed that some of the parameter standard error estimates are comparatively large, in particular the ones corresponding to the latent terms λ_t^X, λ_{it}^{CX}, λ_{jt}^{SX}, λ_{kt}^{LX} and λ_{gt}^{GX}, even when there is complete information. It can be seen that in comparison with the parameter estimates the standard errors are remarkably large and this is indicative of the flat likelihood of the parameters.

Table 6.10: Complete Data Variance of Parameter Estimates

Parameter	Estimate	SE (under complete data)
λ	5.1321	0.07660
λ_i^C	-3.4848	0.4195
λ_j^S	-3.2994	0.3847
λ_k^L	-23.0906	4467.0538
λ_t^X	-2.4349	0.2619
λ_g^G	-49.9015	7260.6983
λ_{it}^{CX}	27.0331	5217.0566
λ_{jt}^{SX}	3.5856	0.4070
λ_{kt}^{LX}	25.4396	4467.0538
λ_{gt}^{GX}	25.8126	5049.7584

In summary, here the EM algorithm manages to fit the latent model, but due to the scarcity of information available for the estimation of these parameters, the variance of the estimates are very large. In other words, the latent model is identified, but there is not enough information available to estimate some of the parameters. The same is observed when trying to obtain the asymptotic variance by using the SEM algorithm for the US Census Dress Rehearsal data. It transpired that the additional dependence term (λ_{ij}^{CS}) and the terms involving the latent variable - λ_t^X, λ_{it}^{CX}, λ_{jt}^{SX} - are weakly identified in the model. After fitting the latent class model $\{CS, SL, CX, SX, LX, GX\}$ the complete data standard error estimates are given in Table 6.11.

It is unfortunate that weak identifiability is a problem that affects analysis of data with latent variables. Evidently identifiability is a property of not just the model, but

[1] An interesting theoretical result is that if the information matrix (which is the inverse of the variance matrix) of the log-likelihood function has eigenvalues smaller than zero then the model is unidentifiable.

also of the data. Formann (2003) states that there are primarily two ways in which a model could be weakly identified; trivially the number of classes could be too many and thus the model is over-parameterized. However, more crucially, the number of classes could be correct but there is not enough data to identify the classes. Garrett and Zeger (2000) showed that simply fitting a Bayesian latent model in the hope of overcoming weak identifiability does not necessarily solve this matter. In actuality, from Gelfand and Sahu (1999) it would seem that Bayesian models can better cope with model non-identifiability than weak identifiability. For a non-identified model it is known a priori that there is no data so the prior fully influences the posterior but for a weakly identified model it is not so straightforward to know how much influence the choice of prior has had on the posterior.

Table 6.11: Complete Data Variance of Parameter Estimates for US Census Example

Parameter	Estimate	SE (under complete data)
λ	5.0365	0.0707
λ_i^C	-1.2611	0.0782
λ_j^S	-2.3217	0.1197
λ_k^L	-1.1802	0.0794
λ_t^X	-37.6689	421.5322
$\lambda_{R_2}^G$	0.0091	0.0872
$\lambda_{O_3}^G$	-0.3031	0.0911
$\lambda_{R_3}^G$	-01903	0.0917
λ_{it}^{CX}	19.1799	299.9527
λ_{jt}^{SX}	36.1513	421.5321
λ_{kt}^{LX}	1.2306	0.2136
$\lambda_{R_2t}^{GX}$	0.3147	0.1676
$\lambda_{O_3t}^{GX}$	0.3099	0.1705
$\lambda_{R_3t}^{GX}$	0.6160	0.1674
λ_{ij}^{CS}	-15.5341	299.9526
λ_{jk}^{SL}	0.4119	0.2080

When there is weak identifiability, the prior will dominate the likelihood and this will inevitably end up with a posterior that is fairly similar to the prior. Garrett and Zeger (2000) also showed that even when the sample size is large the likelihood may still be unable to overcome the prior in the presence of weakly identified parameters. Their solution is to calculate an additional parameter, τ_j that gives the percentage overlap between the prior and posterior distributions for each parameter, j. When τ_j is close to 1, then it means that the parameter is weakly identified. Incidentally, Formann (2003), using a result in Goodman (1974) suggested looking at the rank of the Jacobian matrix, and if the Jacobian has full rank then it follows that all the parameters are locally identifiable, which in turn implies that the model is identifiable. Moreover, it is still possible that a specific model could be 'empirically non-identified', meaning that although the model may be theoretically identifiable the observed data does not allow the effectual estimation of all the parameters. In these circumstances, much more advanced Bayesian models that apply MCMC techniques have been suggested by Gimenez et al. (2008).

6.8 Estimation of the Standard Errors Using the Bootstrap

With the advances in computationally intensive methods, bootstrapping is increasingly being used for the estimation of precision in capture-recapture studies. It was mentioned in Chapter 3 that for data collected by capture-recapture a parametric bootstrap is preferred to a non-parametric bootstrap (Buckland and Garthwaite (1991)). The advantage of the bootstrap in this setting is that although computer intensive it avoids the potential pitfalls of the SEM algorithm, particularly in terms of the calculation of the Jacobian - precisely, the DM matrix. Nonetheless the issue of weak identifiability that makes it difficult to find the variance using the SEM algorithm when there is latentness is still an issue here.

The implementation of the bootstrap carried out here is slightly different from the previous bootstrap in Chapter 5, although the ideas are essentially the same. The objective is to resample the data so as to obtain the bootstrap distribution which in turn gives information about the unknown sampling distribution of interest. The bootstrap makes it possible to draw inferences about the true, but unknown, population based on sample. It was mentioned in Section 3.7.4 that there are two ways of Bootstrapping, the parametric or non-parametric version. The difference between a non-parametric bootstrap and a parametric one is that the resamples are done under a probability model. In essence, a parametric bootstrap requires the choice of an underlying distributional model (in this case a multinomial), while the non-parametric does not.

In the bootstrapping carried out in this chapter, the original data are in the form of a 2x2x2xg contingency table of counts, where g represents the levels of the grouping covariate, G. As such the observed table has cell counts $\{n_{ijkg}\}$, with $\{n_{000g}\}$ being missing. Under the bootstrap, B 'new' contingency tables $\left\{n^{*b}_{ijkg}\right\}$ are generated from the original table, where $b = 1, 2, \cdots, B$. These new contingency tables are the bootstrap resamples. Subsequently, the EM algorithm (detailed above in Section 6.4) is used to fit the model to each of the resamples. This obtains parameter and fitted estimates for each resample. The sample average of the bootstrap resamples when compared to the maximum likelihood estimate, derived under the original observed contingency table, gives an indication of the bias of the maximum likelihood estimate. The bootstrap estimate of the standard error is the standard deviation of the B bootstrap resamples, and this can be taken to be the estimate of the standard error of the maximum likelihood estimate (Efron and Tibshirani, 1993, page 13).

In the bootstrap resampling another issue - which was not encountered previously - that had to be considered was that of *label switching.* Due to the arbitrary nature of the labelling of the latent classes, the results of the EM algorithm fitted to the resampled data is often invariant to the permutations in the labelling of latent classes[2]. Essentially, the likelihood function of the latent model is invariant under both permutations of the latent

[2] The latent variable has two classes but there is no way of knowing in advance how the algorithm splits the data: it could place real enumerations in class 1 and erroneous enumerations in class 2, or vice versa.

classes, i.e. changing the ordering of the latent classes does not change the likelihood value. So to properly draw inferences about the bootstrapped resamples label switching needs to be explicitly addressed, or else there could be a distortion of the statistics of interest.

Appendix B.6 shows the program that was written in SPLUS/R to perform the bootstrapping and has a part that examines the output at convergence of the EM algorithm in order to take account of label switching. As such the estimates are unaffected by changes to the specification of the latent classes. Noticeably, label switching is well known in Markov Chain Monte Carlo (MCMC) methods and there are a number of techniques to deal with them in Bayesian analysis, for example by placing restrictions on the parameters or graphical displays (Garrett and Zeger (2000)).

One more issue involves whether or not the sample truly is representative of the population. In bootstrapping the resamples are assumed to approximate the distribution of the estimator, using the observed sample. In general the bootstrap resamples will approximate the distribution of the unknown estimator, but the bootstrap resamples will exhibit some bias. Ideally, the bias of the mean of the bootstrap resamples, $E\left(\hat{\theta}^{*}\right) - \hat{\theta}$, should be similar to the bias of the maximum likelihood estimate, $E\left(\hat{\theta}\right) - \theta$. Owing to flat observed data likelihood the EM algorithm when applied to the bootstrap resamples will fail to converge to the 'correct' solution in some cases. As such, the bootstrap implemented here will try to solve this by multiplying the observed data by an adjustment factor, f_b. This has the effect of increasing the sample size and producing more precise estimates of variance.

This does suggest that the same idea of inflating the sample size could be extended to the SEM algorithm so as to obtain estimates of the asymptotic variance. Nonetheless, even after multiplying the observed data by 1000, the SEM algorithm was found to still have difficulties in estimating the variance-covariance matrix. It was mentioned previously that there are two ways in which a model may be weakly identified; either due to the actual model or due to the data. The second case, referred to as *empirical weak identifiability* by Garrett and Zeger (2000) can be corrected for by simply increasing the sample size. The fact that the SEM algorithm fails to converge to the asymptotic covariance-variance matrix, even after the sample size increase, suggests that the weak identifiability is due the difficulties in the latent model and not the data.

6.8.1 Derivation of Bootstrap Variances for the Feasibility Study Data

Bootstrapping was implemented to find measures of precision for the Feasibility Study data using the procedure described above. In order to demonstrate the efficacy of the adjustment factor on the bootstrap estimated standard errors, three different values of $f_b = \{1, 100, 1000\}$ were considered. The case when $f_b = 1$ denotes the resamples without any adjustment. It was found during the bootstrapping exercise that the observed population, with sample size $n = 933$, has a data likelihood that is difficult to find precise estimates of the standard error. Changing the stopping criterion[3] to be smaller than the current tolerance (of 10^{-6}) did not seem to make much of a difference, apart from increasing the time to convergence. Therefore, the simpler approach was to multiply the observed population by the adjustment factor, then use these adjusted data as the bootstrap replicates. The EM algorithm was then used to obtain the maximum likelihood estimates for each resample. The estimate of the variance of the maximum likelihood estimate is the sample variance of the bootstrap resamples inflated to take account of the adjustment factor. As will be demonstrated in the following results, the parameter values - apart from the intercept parameter[4] - remain unchanged. This approach was found to be far more computationally efficient and speeds about the processing time; it took roughly a third of the time in comparison to the first approach.

Table 6.12: Feasibility Study - Bootstrap Means and Standard Errors of Parameter Estimates

Parameter	Estimate	Bootstrap Mean	Bootstrap SE	Bootstrap Mean	Bootstrap SE	Bootstrap Mean	Bootstrap SE
		1000 resamples		100 resamples		1000 resamples	
		f_b=1, $n \times f_b = 993$		f_b=1000, $n \times f_b = 933000$		f_b=100, $n \times f_b = 93300$	
λ	2.7731	2.3045	3.5833	9.6916	1.4833	7.4538	1.9781
λ_i^C	0.2633	0.2877	4.1310	0.2604	1.5791	0.1538	2.1957
λ_j^S	-0.4057	-1.5877	5.2473	-0.4797	2.1967	-0.6707	2.8726
λ_k^L	-1.3863	-1.3656	4.1179	-1.4638	2.4563	-1.6767	3.1667
λ_t^X	-0.4070	-4.8981	9.8961	-0.3988	1.6029	-0.4272	1.9719
λ_g^G	-0.0015	0.5110	1.6885	-0.0148	0.7730	-0.0701	0.7731
λ_{it}^{CX}	0.4055	2.5964	8.8070	0.4307	1.7146	0.5406	2.3242
λ_{jt}^{SX}	1.7918	6.6092	9.3679	1.8592	2.2303	2.0455	2.8924
λ_{kt}^{LX}	1.7918	2.4070	5.1817	1.8631	2.4272	2.0623	3.1220
λ_{gt}^{GX}	0.8139	0.6554	2.9550	0.8232	0.7458	0.8709	0.7750
λ_{ij}^{CS}	-0.9808	-2.9343	6.2491	-0.9793	0.1801	-0.9841	0.2399

Table 6.12 gives the estimated standard error of the parameters, and demonstrates the consistency of the bootstrap for the different values of f_b. The bootstrap standard errors presented are the standard deviations based on the resamples multiplied by square root of the adjustment factor. There are two main points that come out of these results. Firstly, the standard errors for $f_b = 100$ and $f_b = 1000$ are roughly similar. Secondly, the estimated standard errors for the non-adjusted bootstrap resamples are rather large when compared to the adjusted resamples. It can also be concluded that the results show that the resampling estimates for $f_b = 100$ are almost equivalent to those for $f_b = 1000$. Therefore, by increasing the sample size to 93300, and then implementing the bootstrap precise estimates of the standard error around the maximum likelihood estimate can be found. However, increasing the sample size to 933000 does not greatly improve the precision. Since it is expected that the variance of an estimator, will decrease proportional to the inverse of the sample size, it would seem that the variance estimates stabilise after $f_b = 100$.

This shows that the choice of the number bootstrapping resamples, B, and adjustment factor, f_b, is based on the considerations of computational cost and not on numerical accuracy. Thus, though the observed data likelihood is flat, multiplying the data by a

[3] The stopping criterion is the point at which it is believed that the parameter estimates are stable and no further improvements can be made to the log-likelihood.

[4] There is a simple relationship between the bootstrap resamples mean and the MLE intercept i.e. $\lambda^{boot} - \log(f_b) = \lambda^{MLE}$.

factor has the effect of 'increasing' the sample size, and as such there is more information available so that the estimators are more precise. By making the likelihood more curved, the curvature can more accurately be estimated, specifically when the observed likelihood is particularly flat.

Table 6.13: Feasibility Study - Bootstrap Means and Standard Errors of Fitted Cell Counts

Cell	Observed Real	Fitted Real	Fitted Real	Observed Erroneous	Fitted Erroneous	Fitted Erroneous
		f_b=1	f_b=100		f_b=1	f_b=100
		$n \times f_b = 933$	$n \times f_b = 93300$		$n \times f_b = 933$	$n \times f_b = 93300$
$\hat{n}_{000g_1}$	**10.66**	**4.6540**	**10.8605**	**15.99**	**27.4076**	**16.1984**
		(3.8236)	(5.5855)		(32.6499)	(24.1540)
n_{100g_1}	21.31	**13.9450**	**21.6804**	21.31	**27.6302**	**21.0014**
		(9.5796)	(10.8963)		(11.4339)	(11.3708)
n_{010g_1}	42.62	**32.5560**	**43.1272**	10.66	**20.8661**	**10.0282**
		(13.3804)	(15.4287)		(16.0532)	(18.3749)
n_{110g_1}	31.97	**26.4039**	**32.3438**	5.33	**11.1818**	**4.8970**
		(10.9945)	(11.1750)		(12.1839)	(12.3517)
n_{001g_1}	15.98	**7.9204**	**16.1849**	3.40	**12.0348**	**2.3304**
		(6.2217)	(6.5990)		(7.6091)	(8.5609)
n_{101g_1}	31.97	**21.3093**	**32.3064**	5.33	**16.4374**	**4.8745**
		(11.7121)	(11.7858)		(12.7878)	(13.3094)
n_{011g_1}	63.94	**53.8428**	**64.3486**	2.66	**12.8907**	**2.3304**
		(17.7570)	(12.2660)		(15.2382)	(7.3384)
n_{111g_1}	47.95	**42.9090**	**48.1892**	1.33	**6.4710**	**11.3845**
		(13.2386)	(10.0086)		(9.8830)	(4.6572)
$\hat{n}_{000g_2}$	**24.01**	**10.9750**	**24.3738**	**16.01**	**43.3443**	**15.9637**
		(8.7127)	(10.7757)		(54.2051)	(25.8579)
n_{100g_2}	48.02	**29.9752**	**48.6340**	21.34	**40.2900**	**20.6970**
		(17.1230)	(20.4004)		(19.8142)	(23.8706)
n_{010g_2}	96.05	**75.0469**	**96.8591**	10.67	**32.5349**	**9.8973**
		(27.8117)	(25.2081)		(27.5222)	(23.5129)
n_{110g_2}	72.04	**60.4069**	**72.5614**	5.33	**16.1604**	**4.8291**
		(20.7153)	(17.6044)		(19.7195)	(14.2923)
n_{001g_2}	36.02	**19.4019**	**36.3386**	4.00	**20.6654**	**3.7036**
		(14.6770)	(12.9724)		(14.7091)	(10.2364)
n_{101g_2}	72.04	**50.3110**	**72.5272**	5.34	**26.4805**	**4.8124**
		(24.9572)	(21.3194)		(23.0648)	(15.6562)
n_{011g_2}	144.07	**124.4133**	**144.2838**	2.67	**21.9917**	**2.2984**
		(31.8458)	(14.2936)		(29.1784)	(8.4268)
n_{111g_2}	108.05	**98.7714**	**108.2244**	1.33	**10.1521**	**1.1251**
		(19.7711)	(11.0000)		(17.0357)	(5.1429)

Table 6.13 gives the resample mean of the fitted cell counts for the different bootstraps with an adjustment factor of $f_b = 100$, and without an adjustment, i.e. $f_b = 1$. Their respective standard errors are also presented and shown in brackets. Again these standard errors are the sample standard deviations of the bootstrap replicates multiplied by the square root of the adjustment factor. The fitted estimates presented are the average of fitted cell counts of the bootstrap resamples divided by the adjustment factor. In addition, since the data has been simulated the actual cell counts are known so have been presented.

Following on from the results in Table 6.12 for the parameter estimates, it can be observed in Table 6.13 that the bootstrap resample means with the adjustment are closer to the actual values, which can be indicative of the non-adjusted bootstrap resample mean being a biased estimator of the population. There is therefore some evidence to suggest that the estimation of the variance may be imprecise at small sample sizes, and it may be appropriate to increase the sample size by an adjustment factor, f_b, and then inflating the sample variance by the adjustment factor to obtain the estimated variance of the parameter estimator.

Furthermore, the standard errors of those that appear in the Third List seem to be larger than those that do not appear in the Third List. A possible reason for this could be simply due to the way in which the Feasibility Study was designed. The probability of being correctly enumerated on the Third List is lower than the Census or Survey, but in contrast the probability of being erroneously enumerated on the Third List is higher than on the Census or Survey. As with any simulation study, the exact nature of the results depends, to a certain extent, on the way in which the data has been simulated. In the next section, when the bootstrap is implemented for the US Census Dress Rehearsal data, these standard errors are much smaller. Thus, it would appear that in this case there may not be sufficient information contained in the observed data to estimate both the latent variable effects and the missing cell counts to a decent degree of accuracy.

6.8.2 Derivation of Bootstrap Variances in Application to the US Census 1990 Dress Rehearsal Data

Measures of accuracy are now presented for the estimates derived under the latent class model fitted to the US Census Dress Rehearsal data discussed in Section 6.6. A similar bootstrapping procedure to the one carried out for the Feasibility Study, in the previous section, was applied here. For this case, the observed sample size, n, is 1013 and the grouping covariate, G, has four levels. However, as previously, B new contingency tables are generated from the original table (shown in Table 5.6). The EM algorithm is next used to fit the latent model so as to obtain maximum likelihood parameter estimates for each of the B resamples. An adjustment factor $f_b = 100$ is also multiplied by the observed sample size to investigate the effect of increasing the observed sample on the variance estimation. The bootstrap estimates of the standard errors presented have been inflated to take this adjustment into account.

Table 6.14 gives the standard errors of the parameter estimates, and Table 6.15 gives the standard errors for the fitted cell counts, for the bootstrap resamples. It can be remembered that when $f_b = 100$ the intercept parameter for the bootstrap resamples has to be adjusted to get the same intercept estimate for the maximum likelihood estimate; here it can be seen that $9.6406 - \log(100)$ is roughly equivalent to the MLE intercept of 5.0365. The first thing to be noticed when comparing the parameter estimates in Table 6.14 to those of the previous application, in Table 6.12, is that the standard errors are much smaller (in magnitude) for this application. Furthermore, the standard errors are also considerably larger for some of the parameters when $f_b = 1$ in comparison with when $f_b = 100$.

Table 6.14: US Census Dress Rehearsal - Bootstrap Means and Standard Errors of Parameter Estimates

Parameter	Estimate	Bootstrap Mean	Bootstrap SE	Bootstrap Mean	Bootstrap SE
		f_b=1		f_b=100	
		$n \times f_b = 1013$		$n \times f_b = 101300$	
λ	5.0365	5.1260	0.4928	9.6406	0.1851
λ_i^C	-1.2611	-1.6853	1.6534	-1.2619	0.1568
λ_j^S	-2.3217	-3.3157	3.6768	-2.2413	0.6161
λ_k^L	-1.1802	-1.5843	0.9048	-1.1814	0.1522
λ_t^X	-37.6689	-35.8061	17.8150	-37.8421	8.5775
$\lambda_{R_2}^G$	0.0091	0.1799	0.5099	0.0171	0.1379
$\lambda_{O_3}^G$	-0.3031	0.0511	0.3223	-0.0402	0.1311
$\lambda_{R_3}^G$	-0.1903	0.02917	0.5126	-0.1774	0.1667
λ_{it}^{CX}	19.1799	19.6351	14.1880	19.5759	9.4517
λ_{jt}^{SX}	36.1513	25.9096	14.7077	35.8299	3.9551
λ_{kt}^{LX}	1.2306	2.7809	11.6891	1.1045	0.8908
$\lambda_{R_2t}^{GX}$	0.3147	0.4325	2.6925	0.3072	0.2010
$\lambda_{O_3t}^{GX}$	0.3099	0.5916	2.2528	0.3384	0.2543
$\lambda_{R_3t}^{GX}$	0.6160	0.6880	2.6139	0.6073	0.2175
λ_{ij}^{CS}	-15.5341	-9.1988	9.1076	-15.5229	0.9906
λ_{jk}^{SL}	0.4119	1.0099	5.4369	0.5436	0.9190

In Table 6.15 the bootstrap resamples without any adjustment (i.e. $f_b = 1$) appear to be more variable, especially the missing cells, $\{n_{000gt}\}$. For the bootstraps with the adjustment, it can be seen that, discounting the missing cells, the standard errors for the n_{011gt} and n_{010gt} cells are comparatively larger than the other cells. Consequently it may be concluded that for those that appear in the Survey but not in the Census, there is insufficient information available from the observed counts to estimate the latent counts to a high degree of precision, even for the increased sample size.

For $f_b = 1$, the mean of the bootstrap resamples are mostly different to those expected under the original sample. In particular for the cells with zero expected counts namely, $\{n_{110g1}, n_{111g1}, n_{000g2}, n_{100g2}, n_{001g2}, n_{101g2}\}$, the bootstrap resample means are non-zero. This does cause some difficulties when it comes to trying to use the results from $f_b = 1$ to find interpretations for the two latent classes. The interpretation, given in Section 6.6, was that there were two classes with everyone who appears in the n_{000}, n_{001}, n_{100} and n_{101} cells belonging in the first latent class. These four cells represent those that were missed by the Survey. Contrastingly, everyone counted by both the Census and Survey - n_{110} and n_{111} - belongs to the second class. The remainder, n_{010}, n_{011}, referring to those missed by the Census but counted by the Survey were shared out amongst the two classes by some mechanism. A much more substantive interpretation was that the two groups denoted whether or not the people were hard or easy to enumerate by the Survey, even after taking into account the post-strata age, race and tenure. Clearly, using the non-adjusted bootstrap results does not make it obvious to find formative descriptions of the two classes.

Table 6.15: US Census Dress Rehearsal - Bootstrap Means and Standard Errors of Fitted Cell Counts

	Estimate	Bootstrap Mean	Bootstrap SE	Bootstrap Mean	Bootstrap SE
		$f_b = 1$		$f_b = 100$	
		O2 Latent Class 1			
n_{000}	155.34	211.10	296.1111	153.78	28.5247
n_{100}	31.00	29.56	8.8524	31.02	5.6083
n_{010}	6.56	7.37	3.3703	7.01	3.7126
n_{110}	0.00	3.32	5.8780	0.00	0.0000
n_{001}	59.00	49.82	20.0759	58.97	7.4725
n_{101}	19.00	15.90	8.6348	18.98	4.1340
n_{011}	10.84	15.98	7.7551	13.20	17.9172
n_{111}	0.00	10.25	24.3602	0.00	0.0000
		R2 Latent Class 1			
n_{000}	155.34	329.28	699.3176	156.44	28.7476
n_{100}	41.00	38.47	9.5737	41.14	6.3668
n_{010}	26.14	29.51	11.9965	28.40	16.9839
n_{110}	0.00	14.34	27.2032	0.00	0.0000
n_{001}	43.00	40.01	8.9281	43.03	6.2047
n_{101}	12.00	10.22	4.4880	12.01	3.5317
n_{011}	5.42	8.54	4.1432	6.85	11.0005
n_{111}	0.00	8.89	17.2917	0.00	0.0000
		O3 Latent Class 1			
n_{000}	149.34	296.4244	223.41	147.73	31.9806
n_{100}	62.00	15.5214	57.22	61.82	7.3581
n_{010}	7.70	4.2891	9.13	8.25	4.9622
n_{110}	0.00	13.5495	7.50	0.00	0.0000
n_{001}	35.00	11.1144	30.63	34.88	5.4532
n_{101}	13.00	5.3468	10.17	12.97	3.5778
n_{011}	4.94	3.8768	7.58	6.16	9.5936
n_{111}	0.00	26.5601	11.54	0.00	0.0000
		R3 Latent Class 1			
n_{000}	127.26	304.96	712.8302	128.79	25.2065
n_{100}	32.00	30.06	6.8402	32.12	5.7929
n_{010}	17.06	19.88	8.64409	18.82	13.7603
n_{110}	0.00	14.53	27.3955	0.00	0.0000
n_{001}	43.00	40.11	9.8910	43.15	6.3419
n_{101}	7.00	5.70	3.1619	7.02	2.7783
n_{011}	5.44	9.11	4.4112	7.11	13.3538
n_{111}	0.00	10.91	21.4719	0.00	0.0000
		O2 Latent Class 2			
n_{000}	0.00	4.86	15.6938	0.00	0.0000
n_{100}	0.00	2.12	7.0831	0.00	0.0000
n_{010}	1.44	0.78	1.8274	1.05	2.9390
n_{110}	13.00	9.66	5.7134	12.95	3.5289
n_{001}	0.00	7.84	16.0781	0.00	0.0000
n_{101}	0.00	3.43	7.0041	0.00	0.0000
n_{011}	8.16	3.93	6.3933	5.89	16.2691
n_{111}	79.00	68.19	26.4283	78.85	8.6280
		R2 Latent Class 2			
n_{000}	0.00	9.37	28.4133	0.00	0.0000
n_{100}	0.00	2.60	6.9878	0.00	0.0000
n_{010}	7.86	4.82	9.9007	5.74	15.6002
n_{110}	69.00	53.91	27.8741	68.82	8.0935
n_{001}	0.00	3.64	7.7682	0.00	0.0000
n_{101}	0.00	1.91	3.9849	0.00	0.0000
n_{011}	5.58	2.14	3.0020	4.13	10.9850
n_{111}	58.00	47.36	18.5258	57.86	7.5132
		O3 Latent Class 2			
n_{000}	0.00	7.91	21.9656	0.00	0.0000
n_{100}	0.00	4.56	13.7982	0.00	0.0000
n_{010}	2.30	1.27	2.5190	1.71	4.4890
n_{110}	36.00	28.48	14.1238	36.10	5.6662
n_{001}	0.00	4.47	9.1707	0.00	0.0000
n_{101}	0.00	2.50	5.1016	0.00	0.0000
n_{011}	5.06	2.30	3.4363	3.82	9.8328
n_{111}	91.00	80.08	29.6119	91.40	8.7271
		R3 Latent Class 2			
n_{000}	0.00	8.22	24.5310	0.00	0.0000
n_{100}	0.00	1.94	5.2517	0.00	0.0000
n_{010}	6.94	3.70	7.1250	5.13	13.7224
n_{110}	69.00	55.18	28.4577	68.96	8.3082
n_{001}	0.00	4.22	9.2250	0.00	0.0000
n_{101}	0.00	1.11	2.3702	0.00	0.0000
n_{011}	7.56	2.82	3.7691	5.73	14.6253
n_{111}	72.00	61.79	24.6635	71.97	8.5193

6.9 The Production of Estimates of the Population for Non-Sampled Areas under Triple System Estimation

There is an additional, fairly important, step in the estimation of population totals that has not been mentioned much during this thesis. In an ecological capture-mark-recapture experiment there are two counts of the population (effectually two 'censuses') and therefore the final population estimate is simply the one derived under dual system estimation. However, in a human census, aside from the financial and time constraints of undertaking two independent censuses, it is not very efficient. Consequently, what often happens is that dual system estimation is applied to the sampled areas in the post-enumeration survey, and it remains to produce population estimates for the non-sampled areas. This is not an easy undertaking, particularly when it is considered that in the 2001 UK Census, the coverage survey sampled roughly 320,000 households and there are in excess of 25 million households in the UK. Visibly, this sample needs to be properly chosen in such a way that it is adequately representative of the entire UK population and this was achieved by use of the Hard-to-Count Index (see Brown et al. (1999) and Chapter 3 of Brown (2000)).

In the 1981 and 1991 censuses, the post-enumeration surveys that were used selected enumeration districts according their expected difficulty to count, and subsequently areas that were considered to be hard to count were over-sampled. The Webber classification (Webber (1977)) was used as the basis of the definition of 'hard-to-count' here and this was arrived at by considering a number of geographical, socio-economic and demographic variables to classify wards and parishes. In 2001 owing to the difficulties in 1981 and 1991, a different hard-to-count classification was derived with the objective of designing a survey such that the selected sample of postcodes could yield an accurate age-sex distribution for each estimation area[5].

It was noted in Chapter 2 that the main shortcoming of the Webber classification was that it classifies small areas based on deprivation. However, the level of underenumeration of an area is not just a factor of its deprivation but also the level of transience. Evidently, people who are highly mobile, attached to multiple households or recent migrants can be expected to be transient and therefore difficult to enumerate. Accordingly, the Hard-to-Count Index was formulated to represent enumeration difficulty based on selected characteristics considered to be importantly related to census underenumeration. The index was constructed from a score calculated using the chosen characteristics for all the enumeration districts in the 1991 Census. These characteristics were

- percentage of heads of households who experienced language difficulty;
- percentage of young people who migrated into the enumeration district in the last year;
- percentage of imputed residents for the enumeration district;

[5] These estimation areas were made of contiguous groups of local authorities with roughly equal population sizes. In 2011 due to the heterogeneity of contiguous local authorities, a more efficient non-contiguous way of grouping similar local authorities is being implemented.

- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented.

The Hard-to-Count Index first used the above characteristics to rank the enumeration districts, then assigned normal scores based on these ranks and finally summed these scores to obtain an overall score each individual enumeration district. Next, the enumeration districts were split into quintiles to create a five level index. All postcodes within enumeration districts were assigned the same hard-to-count index as the enumeration district. The distribution of enumeration districts in each Hard-to-Count stratum was found to be as follows[6]: 7.7% in stratum 1, 19.5% in stratum 2, 27.0% in stratum 3, 28.7% in stratum 4 and 17.2% in stratum 5 (Brown et al. (1999)). Now since every postcode in the country is assigned a hard-to-count score, the Hard-to-Count Index can now be used as the stratification factor in the design of the Census Coverage Survey to select a sample of postcodes.

In the hypothetical case where the Census Coverage Survey is assumed to be perfect such that within the sampled areas a complete coverage is achieved implying there is no underenumeration, then there are a number of standard estimation techniques from survey sampling theory that can be applied to obtain the population totals from the sample. However, in the real census environment, the Census Coverage Survey does miss people and so Chapter 4 of Brown (2000) describes how population totals can be obtained from non-perfect Census Coverage Survey counts. In theory, if a 'complete' population estimate for the sampled areas can be arrived at then it follows that the same standard estimation techniques can still be applied to derive non-sampled area population estimates. The procedure in the 2001 UK Census was to use dual system estimation to adjust the initial Census and subsequent Census Coverage Survey counts for underenumeration. This then yields the corrected population counts for the sampled areas, and ratio estimation was then used to find the population counts for the non-sampled areas.

Ratio estimation is a survey sampling technique used to find the population total from a sample, and the ratio estimator can be derived as follows. Suppose the (unknown) population total is N, and the objective is to find this total through a sample. Define the (known) sample total to be N_s and the (unknown) non-sampled total to be N_r. It follows that the population total can be decomposed as $N = N_s + N_r$.

Now supposing that there is a proportional relationship between two variables $\boldsymbol{Z}$ and $\boldsymbol{Y}$ (i.e. there is a strong correlation between the (Z_i, Y_i) pairs), then Royall (1970) showed that it is justifiable to use the stratification on the known population as an efficient estimator for the unknown population.

So in the 2001 Census, census counts were available for all sampled and non-sampled postcodes, and it is reasonable to assume that the 'true' unknown counts are proportional

[6] This precludes the 28 enumeration districts with a significantly high proportion of 20-34 year old males that were automatically selected into the Census Coverage Survey sample.

to the census counts. Using the same notation, let $\boldsymbol{Z}$ represent the census counts and $\boldsymbol{Y}$ be the 'true' counts. Since there is a linear relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, and an additional assumption is made that the conditional variance of Y_i is proportional to Z_i. Then a plausible way of explaining Y_i given Z_i is via a linear regression model of the form

$$\begin{aligned} E\left[Y_i|Z_i\right] &= \beta Z_i, \\ V\left[Y_i|Z_i\right] &= \sigma^2 Z_i, \\ \mathrm{Cov}\left[Y_i, Y_j|Z_i, Z_j\right] &= 0 \quad \text{for all } i \neq j. \end{aligned} \tag{6.4}$$

β and σ^2 are unknown model parameters and need to be estimated using the available data. If there is a reasonably strong relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, implying that the correlation is close to 1, and if an estimator $\hat{\beta}$ can be found, then it becomes possible to predict Y_i given Z_i by $\hat{\beta}Z_i$. This translates into an estimator of the population total,

$$\hat{N} = N_s + \hat{\beta} \sum_{i \notin s} Z_i. \tag{6.5}$$

Royall (1970) using ordinary least squares techniques showed that the best linear unbiased estimator for β is

$$\hat{\beta} = \frac{\sum_{i \in s} Y_i}{\sum_{i \in s} Z_i}, \tag{6.6}$$

and the corresponding ratio estimator is given by

$$\hat{N} = \left(\frac{\sum_{i \in s} Y_i}{\sum_{i \in s} Z_i} \right) \sum_{i \in \mathcal{U}} Z_i. \tag{6.7}$$

Obviously in the 2001 Census one important feature of the dual system estimation process was the choice of post-strata to ensure homogeneity. Therefore, the version of (6.7) used in the One Number Census as described in Chapter 4 of Brown (2000) was

$$\hat{N}_{RAT} = \sum_{h=1}^{5} \hat{\beta}_h \sum_{e=1}^{D_h} \sum_{m=1}^{M_e} Z_{meh} \tag{6.8}$$

where N_{RAT} is the true population total to be estimated using the ratio model, Z_{meh} is the census count for postcode m from enumeration district e of Hard-to-Count stratum h and D_h is the total number of enumeration districts in Hard-to-Count stratum h, M_e is the total number of postcodes in enumeration district e and $\hat{\beta}_h$ is the least squares estimate of the population ratio of true counts to the census counts, given by

$$\hat{\beta}_h = \frac{\sum_{e=1}^{d_h} \sum_{m=1}^{5} Y_{meh}}{\sum_{e=1}^{d_h} \sum_{m=1}^{5} Z_{meh}}, \tag{6.9}$$

where d_h is the number of enumeration districts sampled in Hard-to-Count stratum h and there are five postcodes sampled from enumeration district e. There are some conditions under which (6.9) is a best linear unbiased estimator of β_h, and these are

$$\begin{aligned} E\left[Y_{meh}|Z_{meh}\right] &= \beta_h Z_{meh}, \\ V\left[Y_{meh}|Z_{meh}\right] &= \sigma_h^2 Z_{meh}, \\ \mathrm{Cov}\left[Y_{meh}, Y_{qpg}|Z_{meh}, Z_{qpg}\right] &= 0 \quad \text{for all } m \neq q. \end{aligned} \tag{6.10}$$

It must be noted that the assumption of zero covariance between postcode counts does not necessarily hold true as a result of the clustered nature of the Census Coverage Survey design - i.e. the postcodes are clustered within enumeration districts. However, in page 72, Brown (2000) using a result from Scott and Holt (1982), showed that this violation does not seriously cause a problem for the unbiased estimation of the total population.

In practice the 'true' counts Y_{meh} are unknown but this is the main reason why dual system estimation is employed. Consequently, the unknown Y_{meh} are replaced by the DSE counts for postcode m, enumeration district e and Hard-to-Count stratum h, $\hat{Y}_{meh}$, which have been shown to be unbiased estimators of the true counts, and the population total becomes

$$\hat{N}_{RAT} = \sum_{h=1}^{5} \frac{\sum_{e=1}^{d_h} \sum_{m=1}^{5} \hat{Y}_{meh}}{\sum_{e=1}^{d_h} \sum_{m=1}^{5} Z_{meh}} Z_h \tag{6.11}$$

where Z_h is the census count across all postcodes in Hard-to-Count stratum h.

When there are now counts from three sources (the Census, Survey and Third List) the proposed methodology is not too different from that detailed above. Since the Survey will still only be occurring in a sub-sample of the population, the same ratio estimation employed in 2001 can be applied to obtain estimates of the population for those non-sampled areas. The only difference is that now $\hat{Y}_{meh}$, the dual system estimate of the population in postcode m, enumeration district e and hard-to-count stratum h is replaced by the triple system estimate, say $\tilde{\hat{Y}}_{meh}$, and the population total is

$$\tilde{\hat{N}}_{RAT} = \sum_{h=1}^{5} \frac{\sum_{e=1}^{d_h} \sum_{m=1}^{5} \tilde{\hat{Y}}_{meh}}{\sum_{e=1}^{d_h} \sum_{m=1}^{5} Z_{meh}} Z_h. \tag{6.12}$$

6.10 Conclusion

The ideas of latent class analysis have been applied in sociological circles for the past half-century or so. However, the basic ideas relied on local independence, which meant that it was assumed that the relationships between the observed manifest variables only existed due to the latent variable, and it becomes possible to divide the population into exhaustive and mutually exclusive latent classes. With the development of log-linear models the local independence assumption can be eased. A further development in the form of the EM algorithm allows locally dependent latent class models to be used in the estimation of the population when there is imperfect capture-recapture data.

Another idea that is heavily relied upon in population estimation, particularly in dual system estimation (and to a lesser extent in triple system estimation), is that of independence. It has been shown that a failure of independence could be firstly due to captures in one list influencing capture in other lists. The dependence that results is what was termed list dependence. The second failure could be a result of the post-stratification mechanism's inability to find properly homogeneous sub-groups, i.e. the stratifying variables fail to prevent the assignment of different types of individuals into the same subgroup; the dependence that results here is what was termed apparent dependence. Clearly, it is easier to have specific measures in place to correct for the second type of dependence than the first. However, for the case when the post-stratification mechanism fails, latent modelling can highlight this failure. The latent class modelling of the US 1990 Census Dress Rehearsal data showed that when the results of log-linear modelling of the capture-recapture contingency table leads to inconclusive estimates of the missing - specifically if there is a difference between the 'saturated' and best-fitting model (as shown in Chapter 5) - the latent model can be beneficial in demonstrating where the post-stratification has not been very adequate. Conversely, when there has been proper post-stratification but there are some erroneous enumerations present in the observed counts, the Feasibility Study showed that latent class analysis is useful in identifying these erroneous counts. Key to the specification of the latent model is the issue of weak identifiability, which occurs when the theoretical conditions for identifiability are met, but the empirical data provides little information about particular parameters.

For weakly identified models variance estimation can be difficult, as they do not have analytic forms particularly for the problem being considered. It was mentioned in Chapter 3 that for capture-recapture methods the information matrix is frequently not easy to invert, and this is the advantage of numerical techniques such as the SEM algorithm. Nonetheless, the presence of a latent variable, combined with the fact that the latent model is, more often than not, weakly identified leads to difficulties in the use of the SEM algorithm. In contrast, bootstrapping techniques produce estimates of precision for any model which parameter estimates can be calculated, and bootstrapping leads to robust confidence interval estimates for capture-recapture models (Buckland and Garthwaite (1991)). In the bootstraps implemented here to obtain estimates of precision around the parameters, it was discovered that owing to the small frequencies in some of cells a large number of bootstrap replicates were needed. In Chapter 6 of their book on bootstrapping, Efron and Tibshirani suggest that a choice of bootstrap resamples in the range of 50 to 200 usually produces reliable standard error estimates. For the bootstraps implemented in Chapter 5, the number of bootstrap resamples needed were 250. However, more resamples were required for the bootstraps undertaken in the presence of a latent variable. In fact it was found that 500 resamples were needed for the US Census application data in order to obtain fairly smooth and stable estimates of the parameters. In contrast, the Feasibility Study data required 1000 bootstrap resamples for stability. This is more in line with the number of replications suggested by Buckland and Garthwaite (1991) who stated that in quantifying precision of estimates in a capture-recapture application 1000 bootstrap resamples should often suffice.

The conclusion is therefore that for imperfect data from an initial Census, a post-enumeration Survey and an Administrative List the latent class formulation of the problem does allow the estimation of the population size. It was shown in Chapter 3 that the log-linear framework of the latent class model can be used to estimate the population size, taking into account of both dependence and capture error. The assumption is that the unobserved heterogeneity is due to the capture error, where this capture error can either be a result of overenumeration or a failure of the post-stratification design. In the 2011 UK Census, the expectation will be that the post-stratification by age-sex group and Hard-to-Count strata does correct for any capture heterogeneity. As such, the application of the triple system estimation described in this thesis will be able to provide an estimate of the population of the UK, adjusted for both underenumeration and overenumeration. The main issue, however, to content with in latent class modelling has been model identifiability, and with it the associated problems of the variance estimation. Therefore, the results of the bootstrapping show that some further thought will be required with regards to whether or not the observed cell counts, after post-stratification, contain enough information (in other words, the data likelihood has some curvature) for the estimation of the variance, in addition to obtaining precise estimates of population sizes.

Chapter 7

Conclusion

7.1 Summary of Main Conclusions

It can be extraordinarily difficult to conduct a census because of a number of reasons. Firstly, the population address frame used to identify households to be given census forms may be incomplete. Secondly, people may be hard to count due to them not being at home when the enumerator visits their household as a consequence of the complexities in modern lifestyle patterns or simply because they do not want to participate in the census, either on malicious grounds or just pure apathy. For this purpose, the census adjustment methodology attempts to address each of these problems and ensure that the final population estimate is as close to the actual population as possible. But admittedly, this is not a straightforward task. The main objective of this thesis, therefore, is to add to existing census adjustment methodology, looking at the areas where there has been perceived failure and come up with a feasible solution. Further, at this period in the run-up to the 2010 round of censuses, the timing of the thesis could not be any more apt.

It has been the aim of the thesis to demonstrate how triple system estimation can be used to obtain population totals, particularly when the data from the three sources is subject to underenumeration and overenumeration. The main motivation of the work lies in the evident inability of dual system estimation to produce unbiased population estimates when there is dependence between the two sources. The (somewhat restrictive) assumption of independence that underpins dual system estimation can now be relaxed under triple system estimation. The three sources that have been considered during this thesis are the initial census enumeration, the post-enumeration survey and an administrative records list. Administrative lists have been widely used in population estimation, and currently several Western European countries, for example the Netherlands, Norway, Finland and Sweden will be relying on them as their main source of population estimates. Further within the UK context administrative data from council tax records, pupil registries, Higher Education student enrolment statistics, National Health patient records etc. have historically been used to quality assure the census figures derived.

Under the proposed methodology, it is assumed that individuals can be cross-classified into a contingency table according to their presence or absence in the initial census count, the post-enumeration survey and the administrative list. During the thesis it has been suggested that health records are the most feasible choice of third list due to the fact that they are (currently) the most encompassing of the whole population unlike other administrative registers, for instance the National Insurance records cover only those above working age. Anyway for such cross-classified data, the log-linear model can be used to model the patterns and associations exhibited by the individuals appearing in the contingency table. It has been shown that a greater range of models that look at the different association structures can be considered, and the best fitting model chosen. Moreover, by treating the correct enumeration status of an individual counted as an unobservable construct the log-linear formulation of latent class analysis can now be employed. The belief is that while the correct enumeration status cannot be directly measured, whether or not an individual appears on the Census, Survey or Third List can be assumed to be imperfect indicators of this. In fact it could be said that the observed capture patterns of individuals are caused by the unobserved, latent, variable that identifies an individual's correct enumeration status. Therefore, it is possible to look at the patterns of interrelationships among the Census, Survey and Third List enumeration histories of individuals to characterize the underlying latent variable of an individual's true enumeration status.

The EM algorithm provides a convenient method for estimation in incomplete data problems, such as the problem encountered here. It has been widely used since Dempster, Laird and Rubin developed it in 1977, and it continues to find uses in a rapidly increasing number of fields, possibly aided by the continued development of different versions and extensions of the algorithm. However, in terms of capture-recapture estimation the use of the EM algorithm has been scarce. This is probably because owing to the work of Cormack, Darroch, Fienberg, amongst others, in log-linear models for incomplete data closed form solutions exist for practically all patterns of associations, and for those that there are no closed form solutions there are simple iterative techniques available. Nonetheless, when the assumptions underlying the capture-recapture estimation models are violated (e.g. there is unobserved heterogeneity), the closed form solutions cannot be used and this is where the EM algorithm comes into its own.

Here the EM algorithm allows the maximisation of the conditional expectation of the complete data log-likelihood since it is easier than trying to maximise the observed data log-likelihood. Precisely, the E and M steps facilitate the computation of the MLEs for incomplete data problems by adapting the techniques used to fit complete data. One unfortunate thing is that while the EM algorithm does provide a simple method for calculating the MLEs it does not automatically give any quantities needed to draw inferences on the calculated MLE parameters, such as the test statistics, and in particular the standard errors. An extension of the EM algorithm, the Supplemented EM (SEM) algorithm, however, does use numerical methods to compute the inverse information matrix, which is the usual estimator of the variance-covariance matrix of the MLEs. Additionally, the common approach for constructing confidence intervals is based on the assumption of asymptotic normality. However, it is well known that in capture-recapture models the small sample distribution of the estimators of the population size are asymmetric and therefore deviate from normality. It follows that normal-based (asymptotic) procedures may frequently produce unreasonable confidence intervals, for instance having lower limits that extend below the number of individuals known to exist or at times being negative. In contrast, the bootstrap and profile likelihood confidence intervals do not have this problem, but are computationally intensive.

Both the SEM algorithm and the Delta method derived confidence intervals are normal-based methods and make use of Taylor Series expansion to linearize the parameter(s) of interest. The Delta method takes the cell probabilities as the parameter(s) of interest and then takes the Taylor's expansion, using the asymptotic normality of the multinomial distribution to estimate the covariance-variance matrix. The parameters are cell probabilities are therefore constrained to lie between $[0, 1]$. In the SEM algorithm, use is made of the logarithm of the cell counts which are linear functions of the parameters of interest, and lie between $[0, \infty]$. The Delta method applies the linearization directly to the parameter, while the SEM applies the linearization to the logged parameter. It is clear that a parameter will converge less slowly to normality than the logarithm of that parameter, and this is reason why the SEM produces more 'realistic' confidence intervals than the Delta Method.

It was also revealed over the course of the thesis that in the capture-recapture models the lack of independence between systems could be due to two things. Firstly, the systems are associated (i.e. there is list dependence) which when there are more than two systems can be easily incorporated into the estimation process. Secondly, across all individuals the inclusion probabilities are not the same (i.e. there is heterogeneity). The second scenario is usually corrected by choosing post-strata that ensure that similar individuals are placed within the same subgroup. In the original log-linear models introduced to estimate population totals from capture-recapture data the assumption was required that there was capture homogeneity and as such the only source of dependence is attributable to list dependence. When there is both list dependence and heterogeneity, the modelling approach has to take this into account or else the estimates of the population will be biased.

Now by characterizing the unobserved heterogeneity as a categorical latent variable, the latent class framework of the log-linear model has shown great promise. Obviously, there are some assumptions, the main one being that the dependence observed among the observed variables is simply due to each of the observed variables' relationship to the latent variable. In essence the latent variable 'explains' the relationships between the observed variables, and is in fact the 'true' source of the dependence observed originally. For some cases, there is some residual dependence not fully explained by the introduction of the latent variable. However, again within a log-linear model these additional dependence terms can be included in the analysis. These log-linear models with a latent variable are not straightforward to fit using iterative methods. Although, the iterative proportional fitting algorithm can be used to fit the simple local dependence latent class model, when there is residual dependence, i.e. the local dependence model, it is much more difficult. So the EM algorithm can be used in order to obtain the MLEs.

In latent class models, a common problem is that of model identifiability. It is suffice to say that a model is identified if there is only one unique solution, on the other hand a non-identified model has more than one solution. When there are three lists, the local independence model is just identified, but only on the proviso that there are no missing cells. Evidently, within a capture-recapture context this raises some potential difficulties with regards to identifiability. However, it has been demonstrated that by the inclusion of a covariate term the ensuing model becomes identifiable. Again there are some conditions for this, in that the covariate needs to be chosen such that the effect of this covariate on the observed variables is only through the latent variable. For triple capture-recapture data both the local independence and local dependence models are now identified, and the model parameters can be estimated. Nonetheless, these models tended to be weakly identified, effectually meaning that the observed data provides little or no information for some of the parameters.

It is apparent that the methods discussed in this thesis are applicable not just to censuses. Wang and Thandrayen (2009) use a similar approach to estimate the number of homeless people in the Australian city of Adelaide's central business district. The authors had data from three agencies that provided services to the homeless. As there is observed and unobserved heterogeneity a mixture model for capture-recapture data (see Böhning et al. (2005)) is fitted to the observed data. Essentially in a mixture model the latent variable is assumed to have a distribution, and when this distribution is discrete then a latent class model results. The latent variable, which accounts for the unobserved heterogeneity can be thought of as being indicative of type of homelessness, for instance long-term or short-term homeless, with the long-term homeless being expected to be easier to count than the short-term.

In their discussion, Wang and Thandrayen found that the log-linear model with a latent variable included to account for unobserved heterogeneity was non-identified, unless some restrictions were placed - they placed equality constraints on some of the conditional probabilities. Furthermore, it was not possible to even fit the local dependence model. However, in an another application, this time using data on the incidence of diabetes in the northern Italian town of Casale Monteferrato (see Bruno et al. (1994)), identifiable latent models can be fitted without recourse to constraints. The main reason for this can be attributed to there now being four lists: diabetic clinic and family physician records (A), hospital admission records (B), computerised insulin prescription records (C) and reimbursement records (D). Here, using the latent variable X to account for the unobserved heterogeneity, it was determined that the best fitting model was the locally dependent model $\{AX, BX, CX, DX, AB, AC, BC, CD\}$, with the first latent class representing patients listed in lists A, B and C only, while the second represented those listed in list D.

It goes to show that the capture-recapture approach holds continuing promise in its applicability to a whole spectrum of fields such as biology, physics, epidemiology and demography. To that end, this thesis has contributed to this area so as to demonstrably aid researchers make informed decisions about the estimation of the population sizes, and the reliability of such derived estimates, in the case when there are data collected from three systems with both observed and unobserved heterogeneity.

7.2 Future Work

The obvious extension of the work undertaken here is to consider using the Bayesian paradigm (such as expanding the models considered in Dellaportas and Forster (1999)), which is ideally suited to missing, and more importantly inadequate, data problems like the capture-recapture problem discussed in this thesis. This is because a Bayesian model will allow the direct combination of prior information with the observed empirical data to provide inferences. In terms of identifiability, and specifically weak identifiability, there is currently work being done using Markov Chain Monte Carlo (MCMC) techniques (for example, Gimenez et al. (2008)) in capture-recapture models. It well known that in Bayesian analysis, non-identifiable models can provide suitable inferences through the examination of the behaviour of the identifiable parameters and the *a priori* information supplied. However, there are two hurdles that need to be negotiated when the model is weakly identified.

The first hurdle is a theoretical one. Due to the data likelihood being flat, the posterior is dominated by the choice of prior which in turn means that the prior has to reflect the prior belief, judgement and uncertainty with a reasonable amount of confidence. When the prior is mis-specified (i.e. placing the wrong prior) the resulting posterior will inevitably be wrong. On the other hand, too informative a prior can lead to a posterior wholly influenced by the prior. This shows the delicate balance needed in the implementation of Bayesian models to capture-recapture data that has latentness. The second - and computational hurdle - is that weak identifiability can result in strong correlations between parameters in the posterior distribution which, subsequently can lead to poor mixing in the MCMC samples and very slow convergence (Gimenez et al. (2008)).

Another extension is in the use of continuous covariates (suggested by Zwane and van der Heijden (2005); Bartolucci and Forcina (2006); Thandrayen and Wang (2009)) to result in a logit-type model. This has the effect of solving the identifiability issues detailed above. The continuous covariates introduced perform the same task as the (discrete) categorical covariates in that they allow for the relaxation of the local independence assumption. However, as in the categorical case, the continuous covariates chosen need to satisfy the assumption that the effect of the manifest variables on this covariate is entirely mediated through the latent variable. Considering that in most properly designed capture-recapture studies care is taken to post-stratify by all covariates known to influence the manifest variables, it is sometimes conceptually difficult to think of a categorical variable not directly included in the post-stratification mechanism but which has an effect on the latent variable. Therefore, in the case when the unobservable heterogeneity is attributable to failure in post-stratification, continuous covariates can be thought of having a limited advantage. On the other hand, continuous covariates become useful when it is known in advance that the post-stratification is fairly accurate such that the unobserved heterogeneity is attributable to erroneous enumerations present in the observed counts. Again, this logit-type model can be formulated in the Bayesian paradigm to make use of prior information.

Finally, for the bootstrap resampling carried out in the thesis, the bootstrap standard errors were used to provide estimates of precision and to give some indication as to the variability of the estimator. Rather than estimating the standard errors, an extension is to use various measures suggested by Efron, for example the confidence intervals. These generally require an increased number of bootstrap resamples. However, confidence intervals can better give the distribution of the parameter than the standard error, in particular for cases where the unknown parameter distribution does not exhibit asymptotic normality. In addition, the bootstrap resampling could be improved further by applying a bias correction as suggested by Efron (1983).

Appendix A

Variances of the Dual and Triple System Estimators (under independence)

A.1 Derivation of the Dual System Estimator Variance under the Delta Method

Recall that the DSE is

$$\hat{N} = n_{11} + n_{10} + n_{01} + \frac{n_{01}n_{10}}{n_{11}} = \frac{n_{1+}n_{+1}}{n_{11}}. \tag{A.1}$$

The Delta method states that the variance of a function $f(x)$ is

$$V\left(f\left(x\right)\right) \simeq \left[\left(\frac{\partial f}{\partial x}\right)^2\right]_E V\left(x\right) \tag{A.2}$$

where $[\,]_E$ is the substitution of the expected values of x appearing inside the brackets after differentiation.

Given that a person is observed, then define the probability of being counted in the census to be p_{cen} and the probability of being counted in the survey to be p_{sur}. Also, suppose that the population total is known, and is N. Further assuming that observed marginal counts are known - i.e. n_{1+} and n_{+1} are fixed - then the number of people counted by both the census and survey, n_{11} can be treated as binomially distributed, with expressions for the mean and variance respectively given by

$$\begin{aligned}&E\left(n_{11}\right) = Np_{cen}p_{sur} = \frac{n_{1+}n_{+1}}{N} \quad \text{and}\\&V\left(n_{11}\right) = Np_{cen}\left(1-p_{cen}\right)p_{sur}\left(1-p_{sur}\right) = \left(\frac{n_{1+}n_{+1}}{N}\right)\left(1-\frac{n_{1+}}{N}\right)\left(1-\frac{n_{+1}}{N}\right)\end{aligned}$$

since $p_{cen} = \frac{n_{1+}}{N}$ and $p_{sur} = \frac{n_{+1}}{N}$.

Consequently, the variance of the population estimate by the dual system estimator is

$$V\left(\hat{N}\right) = V\left(\frac{n_{1+}n_{+1}}{n_{11}}\right) = \left(n_{1+}\right)^2\left(n_{+1}\right)^2 V\left(\frac{1}{n_{11}}\right).$$

The Delta method can now be used, such that

$$\begin{aligned} V\left(\frac{1}{n_{11}}\right) &\simeq \left[\left(\frac{\partial}{\partial n_{11}}\frac{1}{n_{11}}\right)^2\right]_E \times V\left(n_{11}\right) \\ &= \left(-\frac{1}{\left[E\left(n_{11}\right)\right]^2}\right)^2 \times \left(\frac{n_{1+}n_{+1}}{N}\right) \times \left(1-\frac{n_{1+}}{N}\right) \times \left(1-\frac{n_{+1}}{N}\right). \end{aligned}$$

Hence it follows that

$$V\left(\frac{1}{n_{11}}\right) \simeq \frac{N^4}{\left(n_{+1}n_{1+}\right)^4} \times \frac{n_{1+}n_{+1}n_{0+}n_{+0}}{N^3}.$$

So

$$\hat{V}\left(\hat{N}\right) \simeq \frac{\hat{N}n_{+0}n_{0+}}{n_{1+}n_{+1}}.$$

Finally substituting $n_{00} = \frac{n_{01}n_{10}}{n_{11}}$ and $\hat{N} = \frac{n_{1+}n_{+1}}{n_{11}}$ and using the fact that $n_{1+}n_{+1} = n_{01}n_{10} + n_{01}n_{11} + n_{10}n_{11} + (n_{11})^2$, the asymptotic variance calculated under the Delta method is given by

$$\hat{V}\left(\hat{N}\right) \simeq \frac{n_{1+}n_{+1}n_{01}n_{10}}{\left(n_{11}\right)^3}.$$

A.2 Derivation of the Naïve Triple System Estimator Variance under the Delta Method

For data collected from three independent samples of the population, then the probability of being found in the $(i,j,k)^{th}$ cell is

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}. \qquad \text{(A.3)}$$

Supposing that the population size is N, then equation (A.3) can be re-written

$$\frac{\hat{n}_{ijk}}{\hat{N}} = \frac{\hat{n}_{i++}}{\hat{N}}\frac{\hat{n}_{+j+}}{\hat{N}}\frac{\hat{n}_{++k}}{\hat{N}},$$

and so, the triple system estimate of the population size (under independence) is

$$\hat{N} = \sqrt{\frac{\hat{n}_{i++}\hat{n}_{+j+}\hat{n}_{++k}}{\hat{n}_{ijk}}}. \qquad \text{(A.4)}$$

$\hat{N}$ is the maximum likelihood estimate, but does not have a closed form solution when n_{000} is unobserved.

There is another estimator, $\bar{\hat{N}}$ (referred to as the 'naïve' estimator),

$$\bar{\hat{N}} = \sqrt{\frac{n_{1++}n_{+1+}n_{++1}}{n_{111}}} \qquad \text{(A.5)}$$

which however has a closed form solution, for fixed and known n_{1++}, n_{+1+}, n_{++1}.

The Delta method can accordingly be used to find the asymptotic variance of this estimator.

Now, if there is independence between the census, survey and third list, then the cell counts in the corresponding 2x2x2 contingency table are multinomially distributed. So defining p_{cen}, p_{sur} and p_{adm} as the probabilities of being observed in the census, survey and list, such that $p_{cen} = \frac{n_{1++}}{N}$,

$p_{sur} = \frac{n_{+1+}}{N}$ and $p_{adm} = \frac{n_{++1}}{N}$. Assuming that N, n_{1++}, n_{+1+} and n_{++1} are fixed, then the mean and variance of n_{111} are

$$\begin{aligned}
E\left[n_{111}\right] &= N p_{cen} p_{sur} p_{adm} = \frac{n_{1++} n_{+1+} n_{++1}}{N^2} \\
V\left[n_{111}\right] &= N p_{cen}\left(1 - p_{cen}\right) p_{sur}\left(1 - p_{sur}\right) p_{adm}\left(1 - p_{adm}\right) \\
&= N \frac{n_{1++}}{N}\left(1 - \frac{n_{1++}}{N}\right) \frac{n_{+1+}}{N}\left(1 - \frac{n_{+1+}}{N}\right) \frac{n_{++1}}{N}\left(1 - \frac{n_{++1}}{N}\right) \\
&= \frac{n_{+1+} n_{1++} n_{++1}}{N^2}\left(1 - \frac{n_{1++}}{N}\right)\left(1 - \frac{n_{+1+}}{N}\right)\left(1 - \frac{n_{++1}}{N}\right).
\end{aligned}$$

Thus the variance of the estimator $\bar{\hat{N}}$ is

$$\begin{aligned}
V\left(\bar{\hat{N}}\right) &= V\left[\left(\frac{n_{1++} n_{+1+} n_{++1}}{n_{111}}\right)^{\frac{1}{2}}\right] \\
&= n_{1++} n_{+1+} n_{++1} V\left[\left(\frac{1}{n_{111}}\right)^{\frac{1}{2}}\right].
\end{aligned}$$

Under the Delta method

$$\begin{aligned}
V\left(\frac{1}{\sqrt{n_{111}}}\right) &\simeq \left[\left(\frac{\partial}{\partial n_{111}} \frac{1}{\sqrt{n_{111}}}\right)^2\right]_E V\left[n_{111}\right] \\
&= \{-\frac{1}{2}\left(E\left[n_{111}\right]\right)^{-\frac{3}{2}}\}^2 \times \frac{n_{1++} n_{+1+} n_{++1}}{N^2}\left(\frac{N - n_{1++}}{N}\right)\left(\frac{N - n_{+1+}}{N}\right)\left(\frac{N - n_{++1}}{N}\right).
\end{aligned}$$

Substituting the expression for the expectation of n_{111}, then

$$\begin{aligned}
V\left(\frac{1}{\sqrt{n_{111}}}\right) &\simeq \frac{1}{4} \frac{n_{1++} n_{+1+} n_{++1}}{N^5} \times \left(N - n_{1++}\right)\left(N - n_{+1+}\right)\left(N - n_{++1}\right) \times \left[\frac{n_{1++} n_{+1+} n_{++1}}{N^2}\right]^{-3} \\
&= \frac{1}{4} \frac{N}{\left(n_{1++} n_{+1+} n_{++1}\right)^2} \times \left(N - n_{1++}\right)\left(N - n_{+1+}\right)\left(N - n_{++1}\right).
\end{aligned}$$

Therefore,

$$\hat{V}\left(\bar{\hat{N}}\right) \simeq \frac{1}{4} \frac{\bar{\hat{N}}\left(\bar{\hat{N}} - n_{1++}\right)\left(\bar{\hat{N}} - n_{+1+}\right)\left(\bar{\hat{N}} - n_{++1}\right)}{n_{1++} n_{+1+} n_{++1}}.$$

Since $\bar{\hat{N}} = \sqrt{\frac{n_{1++} n_{+1+} n_{++1}}{n_{111}}}$, the asymptotic variance of the ‘naïve’ independence triple system estimator is given by

$$\hat{V}\left(\bar{\hat{N}}\right) \simeq \frac{\left(\sqrt{\frac{n_{1++} n_{+1+} n_{++1}}{n_{111}}} - n_{1++}\right)\left(\sqrt{\frac{n_{1++} n_{+1+} n_{++1}}{n_{111}}} - n_{+1+}\right)\left(\sqrt{\frac{n_{1++} n_{+1+} n_{++1}}{n_{111}}} - n_{++1}\right)}{4\sqrt{n_{111}\left(n_{1++} n_{+1+} n_{++1}\right)}}.$$

Appendix B

SPLUS/R Programs

B.1 EM algorithm - no erroneous enumerations

```
EM.sim <-function(tol, data, eqn)
{
r <- length(levels(data[, 1]))
data$em.data <- data$count
data$em.data[c(1:length(data[, 1]))[is.na(data[, 4])]] <- 0
model <- glm(eqn, data = data, family = poisson)
est <- 1
est <- cbind(est,model$coef)
fit <- model$fitted
data$em.data[c(1:length(data[, 1]))[is.na(data[, 4])]]
                <- c(fit)[c(1:length(data[, 1]))[is.na(data[, 4])]]
i <- 2
while(any(c(abs(est[, i] - est[, i - 1])) > tol,na.rm=T))
    {
model <- glm(eqn, data = data, family = poisson)
est <- cbind(est,model$coef)
fit <- model$fitted
data$em.data[c(1:length(data[, 1]))[is.na(data[, 4])]]
                    <- c(fit)[c(1:length(data[, 1]))[is.na(data[, 4])]]
i <- i + 1
}
est<<-est
fit<<-fit
cat("Converged in", i ,"steps", fill=T)
round(array(data$em.data, c(r, r, r)), 2)
lik<<-sum(((data$count[-c(1)]-fit[-c(1)])^2)/fit[-c(1)])
loglik<<- 2*(sum((data$count[-c(1)]) * (log((data$count[-c(1)])/(fit[-c(1)])))))
resid <<- data$count[-c(1)] - fit[-c(1)]
adj.resid <<- resid / stdev(resid)
```

```
    cat("missing =", round(fit[1], 3), fill=T)
cat("X^2 likelihood statistic =", round(lik, 3), fill=T)
cat("G^2 likelihood statistic =", round(loglik, 4), fill=T)
}
```

B.2 SEM algorithm

```
# Program adapted from
# Robert Gray’s Advanced Statistical Computing Course Notes
# University of Wisconsin - Madison, Dept of Statistics
# http://www.stat.wisc.edu/~mchung/teaching/stat471/stat_computing.pdf
# Special Thanks to Guy Abel, Division of Social Statistics,
# University of Southampton

# save the original data and replace NA’s
data$y<-data$count
data$y[is.na(data$y)] <- 0
# run a basic model to get model.matrix
options(contrasts = c("contr.treatment", "contr.poly"))
model1<-glm(y~census+survey+admin+group,poisson, data = data)
# EM function to compute one iteration,
# given model, data and some intial beta estimate
em<-function(beta0, model, data)
{
#E step
fit<-exp(model.matrix(model)%*%beta0)
data$y[is.na(data$count)] <- c(fit)[is.na(data$count)]
#M step
m<-glm.fit(model.matrix(model),data$y, family=poisson())
beta<-m$coef
list(beta=beta,fit=data$y)
}
# some intial beta’s and intial error
beta <-rep(1,length(model1$coef))
err <- 10
i <- 0
# run the EM function until convergence
while(err > 1e-010)
{
i <- i + 1
u <- em(beta, model1, data)
err <- max(abs(u$beta - beta))
beta<-u$beta
fit<-u$fit
```

```
print(c(i, err, beta))
}
## this shows that the 'new' EM algorithm works
## converged mle (this is important)
beta
cbind(data,EM.fit=fit)
## the next part is to do the jacobian matrix
## remember that the jacobian is the matrix of
# all first order partial derivatives
## the jacobian is used to find the score function
## (i.e. the second order derivatives)
# function to compute one iteration of jacobian matrix
## i.e. function computes the jacobian of the EM algorithm mapping
jac <- function(b, beta, model, data)
{
#store mle of b for DM
us <- em(b,model,data)
#intialize
u <- em(beta,model,data)
beta <- u$beta
dm <- matrix(0, length(beta), length(beta))
for(i in 1:length(beta)) {
#sequentially replace each mle beta with starting beta
bb <- b
bb[i] <- beta[i]
#run one iteration of EM with altered betas
u <- em(bb,model,data)
#fill in the relevent DM row with rate of change
dm[i,  ] <- c(u$beta - us$beta)/(bb[i] - b[i])
}
print(dm)
list(dm = dm, beta = beta)
}
# intial values used in EM
bb<-list(beta=rep(1,length(model1$coef)))
# run dm to see rate of change from first iterate to second
dm<-jac(beta,bb$beta,model1,data)$dm
# do this iteratively until DM looks stable
err <- 10
i <- 0
while(err > 1e-005)
{
i <- i + 1
bb <- em(bb$beta,model1,data)
w <- jac(beta,bb$beta,model1,data)
err <- max(abs(c(w$dm-dm)))
dm<-w$dm
```

```
print(i)
}
# compute conditional expectation of complete data information
# refit model to get (asymptotic) covariance matrix
# using the fitted instead of observed data
data<-cbind(data,EM.fit=fit)
model1.f<-glm(EM.fit~census+survey+admin+group,poisson, data)
## this gets out the estimates of the complete data covariance matrix
cov.com <- summary(model1.f)$cov.unscaled
## compute DM, derivative of mapping (i.e. missing information)
dm2 <- -w$dm
## remember want (1-DM), so
diag(dm2) <- diag(dm2)+1
## using Orchard and Woodbury principle
## observed information = complete information - missing information
## also Dempster, Laird and Rubin showed that
## observed-data variance = complete data variance + increase in
## in variance due to missing data
## and so dV = DM ((1-DM)^-1) ((Vcom)^-1)
delta.v <- cov.com %*% w$dm %*% solve(dm2)
## the variance is given by the diagonal elements, with the observed
## covariance added
est.var <- diag(cov.com+delta.v)
## print out the parameter estimates and the variance
beta
est.var
```

B.3 EM algorithm - with erroneous enumerations

```
EM.latent.miss <- function(tol=1e-10, n.ijkg)
{
## set the initial starting values for the missing cells
n.ijkg[is.na(n.ijkg)] <- 1
## Haberman starting values
## These are needed to ensure that the convergence is at the ‘right’ solution
init <- array(c(82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262),
dim=c(2,2,2,2,2))
data.lat <- init
dimnames(data.lat) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"),
                            c("Group A", "Group B"), c("Erroneous", "Real"))
data.lat <- data.frame(expand.grid(census = dimnames(data.lat)[[1]],
    survey = dimnames(data.lat)[[2]], admin = dimnames(data.lat)[[3]],
    group=dimnames(data.lat)[[4]], latent = dimnames(data.lat)[[5]]),
    count.lat = c(data.lat))
r <- length(levels(data.lat[, 1]))
data.lat$em <- data.lat$count.lat
## model - with the XG interaction
eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent
                             + survey:latent + admin:latent + group:latent
n.est <- 1
n.data <- 1
j <- 1
repeat {
model <- glm(eqn, data = data.lat, family = poisson)
mu.ijkgt <- array(c(model$fitted), dim = c(2, 2, 2, 2, 2))
mu.ijkg <- apply(mu.ijkgt, c(1, 2, 3, 4), sum)
#estimate the missing cells
n.ijkg[1]  <- mu.ijkgt[1] + mu.ijkgt[17]
n.ijkg[9]  <- mu.ijkgt[9] + mu.ijkgt[25]
n.ijkg <- n.ijkg
weight <- array(n.ijkg/mu.ijkg, dim = c(2, 2, 2, 2, 2))
n.ijkgt <- weight * mu.ijkgt
data.lat$em <- c(n.ijkgt)
n.est <- cbind(n.est, model$coef)
n.data <- cbind(n.data, c(n.ijkgt))
if ((all(abs(n.data[,j+1] - n.data[,j]) < tol, na.rm=T))
&& (all(abs(n.est[,j+1] - n.est[,j]) < tol, na.rm=T)))
break
j <- j + 1
}
n.data <<- n.data
cat("Converged in", j, "steps", fill = T)
cat ("estimate of latent classes", fill=T)
print(c(round(array(data.lat$em, c(r, r, r, r, r)), 2)))
#check that the observed cells are unchanged, at end of iterations
cat ("estimate of observed cells", fill=T)
print(c(round(apply(array(c(data.lat$em), dim=c(2,2,2,2,2)), c(1,2,3,4), sum),2)))
}
```

B.4 US 1990 Census Data - fitting a latent class model

```
############################################################################
# Using the US Census Rehearsal Data to investigate the possibility of     #
## some unobserved heterogeneity (latentness) in the data structure        #
############################################################################

US.EM.latent.miss <- function(tol=1e-5)
{
## grouped dataframe
n.ijkg <- array(c(NA,31,8,13,59,19,19,79,
      NA,41,34,69,43,12,11,58,
      NA,62,10,36,35,13,10,91,
          NA,32,24,69,43,7,13,72), dim=c(2,2,2,4))
n.ijkg[is.na(n.ijkg)] <- 1
## Haberman starting values
init <- array(c(82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262),
dim=c(2,2,2,4,2))
data.lat <- init
dimnames(data.lat) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"),
c("Group A", "Group B","Group C", "Group D"), c("Erroneous", "Real"))
data.lat <- data.frame(expand.grid(census = dimnames(data.lat)[[1]],
survey = dimnames(data.lat)[[2]], admin = dimnames(data.lat)[[3]],
    group=dimnames(data.lat)[[4]], latent = dimnames(data.lat)[[5]]),
    count.lat = c(data.lat))
r <- length(levels(data.lat[, 1]))
    ## M - Step : Specify the model
data.lat$em <- data.lat$count.lat
## model 1 - local independence
#eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent
# model 2 - local dependence - CS interaction
#eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + census:survey
#model 3 - local dependence - SL interaction
#eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + survey:admin
## model 4 - local dependence - CS, SL interactions
# eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + census:survey + survey:admin
# model 5 - local dependence with CSG interaction
# eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + census:survey + census:survey:group
## model 6 - local dependence with SLG interaction
#eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + survey:admin + survey:admin:group
## model 7 - local dependence with CSG, SLG interaction
eqn <- data.lat$em ~ census + survey + admin + group + latent + census:latent +
survey:latent + admin:latent + group:latent + survey:admin + survey:admin:group
n.est <- 1
n.data <- 1
j <- 1
repeat {
model <- glm(eqn, data = data.lat, family = poisson)
mu.ijkgt <- array(c(model$fitted), dim = c(2, 2, 2, 4, 2))
mu.ijkg <- apply(mu.ijkgt, c(1, 2, 3, 4), sum)
        # E - step: Estimate missing cells
n.ijkg[1]  <- mu.ijkgt[1] + mu.ijkgt[33]
n.ijkg[9]  <- mu.ijkgt[9] + mu.ijkgt[41]
n.ijkg[17]  <- mu.ijkgt[17] + mu.ijkgt[49]
n.ijkg[25]  <- mu.ijkgt[25] + mu.ijkgt[57]
weight <- array(n.ijkg/mu.ijkg, dim = c(2, 2, 2, 4, 2))
n.ijkgt <- weight * mu.ijkgt
data.lat$em <- c(n.ijkgt)
n.est <- cbind(n.est, model$coef)
n.data <- cbind(n.data, c(n.ijkgt))
if ((all(abs(n.data[,j+1] - n.data[,j]) < tol, na.rm=T)) &&
        (all(abs(n.est[,j+1] - n.est[,j]) < tol, na.rm=T)))
break
j <- j + 1
}
n.data <<- n.data
cat("Converged in", j, "steps", fill = T)
cat ("estimate of latent classes", fill=T)
print(c(round(array(data.lat$em, c(r, r, r, 4, r)), 4)))
cat ("estimate of observed cells", fill=T)
print(c(round(apply(array(c(data.lat$em), dim=c(2,2,2,4,2)), c(1,2,3,4),sum),2)))
}
```

B.5 Parametric Bootstrap - no erroneous enumerations

```
################################################
# Bootstrapping of US Application Data         #
## Without the latent variable                 #
################################################

n.O2 <- c(31 , 8 , 13 , 59 , 19 , 19 , 79)
n.R2 <- c(41 , 34 , 69 , 43 , 12 , 11 , 58)
n.O3 <- c(62 , 10 , 36 , 35 , 13 , 10 , 91)
n.R3 <- c(32 , 24 , 69 ,43 , 7 , 13 , 72)

## Step 1 - Resampling
## where pi.ijk is the sampling probability based on n.O2, n.O3, n.R2 or n.R3
boot.EM <- function(b,n,pi.ijk)
{
data.boot <- matrix(sample(1:8, n*b, prob=pi.ijk, replace=T),byrow=F, nrow=b)
X.tab <- apply(data.boot, 1, nbins=8, tabulate)
return(X.tab)
}

data.nijk.US <- boot.EM()

# Step 2
### Next need to make sure that the n000 cell is missing
em.dat.na <- function(data)
{
data[c(1)] <- NA
data
}

## Step 3
### Final Data set with missing cells replaced with  NAs
data.n.ijk <- apply(data.nijk.US, 2, em.dat.na)

## Step 4: Implementation of the EM algorithm

## Part 1
em<-function(beta, model, data)
{
#E step
fit<-exp(model.matrix(model)%*%beta)
data$y[is.na(data$count)] <- c(fit)[is.na(data$count)]
#M step
m<-glm.fit(model.matrix(model),data$y, family=poisson())
beta<-m$coef
list(beta=beta,fit=data$y)
}

## Part 2
em.full<-function(beta,model,data,tol)
{
beta.init <-rep(1,length(model$coef))
beta <- beta.init
i<-0
err<-tol*2
storebeta<-beta
while(err > tol) {
i <- i + 1
u <- em(beta, model, data)
err <- max(abs(u$beta - beta))
beta<-u$beta
fit <- u$fit
storebeta<-cbind(storebeta,beta)
}
return(list(i=i, beta=u$beta, fit=u$fit, storebeta=storebeta))
}

## Part 3
EM.US.boot <- function(dat)
{
dat <- array(c(dat), dim=c(2,2,2))
dimnames(dat)<-list(c("no","yes"),c("no","yes"),c("no","yes"))
data<-data.frame(expand.grid(admin=dimnames(dat)[[1]], survey=dimnames(dat)[[2]],
census=dimnames(dat)[[3]]), count=c(dat))
data$y<-data$count
data$y[is.na(data$y)] <- 0
model.EM<-glm(y˜census+survey+admin+census:survey+survey:admin,poisson, data = data)
options(contrasts = c("contr.treatment", "contr.poly"))
EM.imp<-em.full(model=model.EM,data=data,tol=1e-5)
return(list(beta=EM.imp$beta,fit=EM.imp$fit))
}

# Step 5: The bootstrap procedure
### The best fitting model here is the one with six terms
store.EM.beta <-matrix(NA,b,6)
store.EM.fit <-matrix(NA,b,8)

for(i in 1:b){
temp<-EM.US.boot(data.n.ijk[,i])
store.EM.fit[i,] <- temp$fit; store.EM.beta[i,] <- temp$beta}

store.EM.beta
store.EM.fit
```

```
## means
apply(store.EM.beta, 2, mean)
apply(store.EM.fit, 2, mean)

## standard errors
sqrt(apply(store.EM.beta, 2, var))
sqrt(apply(store.EM.fit, 2, var))
```

B.6 Parametric Bootstrap - with erroneous enumerations

```
# Step 1
## Produce b bootstrap samples from the 32 cells (n.ijkgt)
## n = total observed and b = bootstrap resamples, pi = cell probabilities
boot.EM <- function(b,n,pi)
{
data.boot <- matrix(sample(1:32, n*b, prob=pi, replace=T),byrow=F, nrow=b)
X.tab <- apply(data.boot, 1, nbins=32, tabulate)
return(X.tab)
}
data.EM <- boot.EM()

## Step 2
## Create the 16 cells (n.ijkg)
em.dat.fn <- function(data)
{
apply(array(c(data), dim=c(2,2,2,2,2)), c(1,2,3,4), sum)
}

## Step 3
## this produces the 16xb cells
data.n.full <- apply(data.EM, 2, em.dat.fn)

## Step 4
### Next need to remove the 1st and 9th cells
em.dat.na <- function(data)
{
data[c(1,9)] <- NA
data
}

## Step 5
### Final data set with missing cells replaced with NAs
data.n.ijkg <- apply(data.n.full, 2, em.dat.na)
```

```
## Step 6a - Implement the bootstrap
## for b bootstrap resamples and dim(beta) is the number of model parameters
### The bootstrapped data is split into two groups to prevent mixing
### EM.latent.boot is the function that performs the EM algorithm on the observed counts
### So the following function takes the resampled data and applies the EM algorithm
### Each resample outputs beta and fitted estimates which are stored

## Step 6a: this is the EM algorithm for one iteration
em <- function(beta0, model=model.latent, data.lat, n.ijkg)
{
n.ijkg[is.na(n.ijkg)] <- 1
#E steps
## # E1 step
fit <- exp(model.matrix(model)%*%beta0)
mu.ijkgt <- array(c(fit), dim=c(2,2,2,2,2))
## # E2 step
n.ijkg[1]  <- mu.ijkgt[1] + mu.ijkgt[17]
n.ijkg[9]  <- mu.ijkgt[9] + mu.ijkgt[25]
mu.ijkg <- apply(mu.ijkgt, c(1,2,3,4), sum)
weight <- array(n.ijkg/mu.ijkg, c(2,2,2,2,2))
n.ijkgt <- weight * mu.ijkgt
data.lat$y <- c(n.ijkgt)
#M step
m <- glm.fit(model.matrix(model), data.lat$y, family=poisson())
beta <- m$coef
list(beta=beta, fit=data.lat$y)
}

## Step 6b: this runs until convergence
em.full<-function(beta,model=model.latent,data,n.ijkg,tol)
{
n.ijkg[is.na(n.ijkg)] <- 1
i<-0
err<-tol*2
storebeta<-beta
while(err > tol) {
i <- i + 1
u <- em(beta, model.latent,data.lat, n.ijkg)
err <- max(abs(u$beta - beta))
beta<-u$beta
fit <- u$fit
storebeta<-cbind(storebeta,beta)
}
return(list(i=i, err=err, beta=u$beta, fit=u$fit, lik=u$lik, storebeta=storebeta))
}
```

```
## Step 6c - Finally bring it all together
### The following function runs the EM algorithm to produce parameter
### and cell estimates given a 2x2x2x2x2 observed table
#### Has an inbuilt procedure to estimate missing n000g1 and n000g2 cells

EM.latent.boot <- function(data)
{
n.ijkg <- data
init <- array(c(82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
        82.074, 11.107, 11.107, 1.503, 11.107, 1.503, 1.503, 0.203,
0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262,
        0.196, 1.440, 1.478, 10.885, 1.477, 10.885, 11.166, 82.262), dim=c(2,2,2,2,2))
data.lat <- init
dimnames(data.lat) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"),
c("Group A", "Group B"), c("Real", "Erroneous"))
data.lat <- data.frame(expand.grid(census = dimnames(data.lat)[[1]],
survey = dimnames(data.lat)[[2]], admin = dimnames(data.lat)[[3]],
group=dimnames(data.lat)[[4]], latent = dimnames(data.lat)[[5]]),
count.lat = c(data.lat))
data.lat$y <- data.lat$count.lat
#run a basic model to get model.matrix
## Local Dependence Model with CS interaction
model.latent<-glm(y˜census+survey+admin+latent+group+census:latent+survey:latent
+admin:latent+latent:group + census:survey, poisson, data = data.lat)

# Need suitable choice initial beta values
## under local dependence
model.init<-glm(count.lat˜census+survey+admin+latent+group+census:latent
+survey:latent+admin:latent+latent:group + census:survey, poisson, data = data.lat)

beta <- model.init$coef
fit <- exp(model.matrix(model.init)%*%beta)
EM.imp<-em.full(model.init$coef,model=model.latent,data.lat,n.ijkg,1e-05)
## maybe change the convergence criterion to 1e-3??
return(list(beta=EM.imp$beta,fit=EM.imp$fit))
}

# Final Step: The bootstrap procedure

store.EM.beta.1 <-matrix(NA,b,dim(beta))
store.EM.beta.2 <-matrix(NA,b,dim(beta))
store.EM.fit.1 <-matrix(NA,b,32)
store.EM.fit.2 <-matrix(NA,b,32)
```

```
## note that the data switching looks at the observed real counts
## and b.switch is chosen based on the a posteriori classification probabilities

for(i in 1:b)
{
temp<-EM.latent.boot(data.n.ijkg[,i])
if(sum(temp$fit[c(2:8,10:16)]) > b.switch)
{
store.EM.fit.1[i,] <- temp$fit; store.EM.beta.1[i,] <- temp$beta
}
if(sum(temp$fit[c(2:8,10:16)]) <= b.switch)
{
store.EM.fit.2[i,] <- temp$fit; store.EM.beta.2[i,] <- temp$beta
}
}

#### To take care of label switching, need to run this part
EM.switch <- function(data)
{
data[!is.na(data)]
}

### To get the beta parameters and the fitted values
apply(store.EM.beta.1, 2, EM.switch)
apply(store.EM.beta.2, 2, EM.switch)
apply(store.EM.fit.1, 2, EM.switch)
apply(store.EM.fit.2, 2, EM.switch)
```

Bibliography

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics 50*(2), 494–500.

Agresti, A. (2002). *Categorical Data Analysis.* Second Edition, John Wiley & Sons.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics 46*(3), 623–35.

Alho, J. (1994). Analysis of Sampled Based Capture-Recapture Experiments. *Journal of Official Statistics 10*, 245–245.

Alho, J., M. Mulry, K. Wurdeman, and J. Kim (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association 88*(423), 1–130.

Bartlett, M. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement 2*, 248–52.

Bartolucci, F. and A. Forcina (2006). A class of latent marginal models for capture-recapture data with continuous covariates. *Journal of the American Statistical Association 101*(474), 786–794.

Bell, W. (1993). Using Information from Demographic Analysis in Post-Enumeration Survey Estimation. *Journal of the American Statistical Association 88*(423), 1106–1118.

Biemer, P., G. Brown, D. Judson, and C. Wiesen (2001a). Triple System Estimation with Erroneous Enumerations in the Administrative Records List. *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Biemer, P., H. Woltmann, D. Raglin, and J. Hill (2001b). Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis. *Journal of Official Statistics 17*(1), 129–148.

Birch, M. (1963). Maximum likelihood in threeway contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 25*, 220–223.

Bishop, Y., S. Fienberg, P. Holland, J. Richard, and F. Mosteller (1975). *Discrete multivariate analysis: theory and practice.* MIT Press, Cambridge, Massachusetts.

Böhning, D., E. Dietz, R. Kuhnert, and D. Schön (2005). Mixture models for capture-recapture count data. *Statistical Methods and Applications 14*(1), 29–43.

Böhning, D. and D. Schön (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(4), 721–737.

Britton, M. and F. Birch (1985). *1981 Census Post-enumeration Survey: An Enquiry Into the Coverage and Quality of the 1981 Census in England and Wales.* Stationery Office Books.

Brown, G., P. Biemer, and D. Judson (2006). Estimating Erroneous Enumeration in the US Decennial Census using Four Lists. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Brown, J. (2000). *Design of a Census Coverage Survey and its Use in the Estimation and Adjustment of Census Underenumeration: A Contribution Towards Creating a One-Number Census in the UK in 2001.* Ph. D. thesis, University of Southampton.

Brown, J., O. Abbott, and I. Diamond (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(4), 883–902.

Brown, J., I. Diamond, R. Chambers, L. Buckner, and A. Teague (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(2), 247–267.

Bruno, G., E. LaPorte, F. Merletti, A. Biggeri, D. McCarthy, and G. Pagano (1994). National diabetes programs: Application of capture-recapture to count diabetes? *Diabetes Care 17*(6), 548–556.

Buckland, S. and P. Garthwaite (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics 47*(1), 255–268.

Castledine, B. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika 68*(1), 197–210.

Chandrasekar, C. and W. Deming (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association 44*(245), 101–115.

Chapman, D. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics 1*, 131–160.

Citro, C., D. Cork, and J. Norwood (2004). *The 2000 Census: Counting Under Adversity.* National Academies Press. National Research Council (US) - Panel to Review the 2000 Census.

Cochran, W. (1977). *Sampling Techniques.* Third Edition, John Wiley & Sons.

Cohen, M. and B. King (2000). *Designing the 2010 Census: First Interim Report.* National Academies Press.

Congdon, P. (2005). *Bayesian Statistical Modelling.* Second Edition, John Wiley & Sons.

Cormack, R. (1989). Log-linear models for capture-recapture. *Biometrics 45*(2), 395–413.

Cormack, R. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics 48*(2), 567–576.

Cormack, R. M. (1972). The logic of capture-recapture estimates. *Biometrics 28*(2), 337–343.

Coull, B. and A. Agresti (1999). The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies. *Biometrics 55*(1), 294–301.

Cox, D. and D. Hinkley (1974). *Theoretical Statistics.* Chapman & Hall.

Darroch, J. (1958). The Multiple-Recapture Census I. Estimation of a Closed Population. *Biometrika 45*(3-4), 343–359.

Darroch, J. (1962). Interactions in multi-factor contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 24*, 251–263.

Darroch, J., S. Fienberg, G. Glonek, and B. Junker (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association 88*(423), 1137–1148.

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1–31.

Dellaportas, P. and J. Forster (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika 86*(3), 615–633.

Deming, W. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*(4), 427–444.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 39*(1), 1–38.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association 78*(382), 316–331.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap.* Chapman & Hall/CRC.

El-Khorazaty, M., P. Imrey, G. Koch, and H. Wells (1977). Estimating the total number of events with data from multiple-record systems: a review of methodological strategies. *International Statistical Review 45*(2), 129–157.

Ericksen, E., J. Kadane, B. Bailar, R. Fay, I. Fellegi, M. Hansen, P. Hauser, J. Passel, S. Preston, and J. Rolph (1985). Estimating the population in a census year: 1980 and Beyond. *Journal of the American Statistical Association 80*(389), 98–131.

Ericksen, E., J. Kadane, and J. Tukey (1989). Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association 84*(408), 927–44.

Evans, M., Z. Gilula, and I. Guttman (1989). Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika 76*(3), 557–563.

Evans, M., H. Kim, and T. O'Brien (1996). An Application of Profile-Likelihood Based Confidence Interval to Capture: Recapture Estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, 131–140.

Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika 59*(3), 591–603.

Fienberg, S. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey methodology 18*(1), 143–154.

Formann, A. (2003). Latent Class Model Diagnosis from a Frequentist Point of View. *Biometrics 59*(1), 189–196.

Garrett, E. and S. Zeger (2000). Latent Class Model Diagnosis. *Biometrics 56*(4), 1055–1067.

Gazey, W. and M. Staley (1986). Population estimation from mark-recapture experiments using a sequential Bayes algorithm. *Ecology 67*(4), 941–951.

Gelfand, A. and S. Sahu (1999). Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. *Journal of the American Statistical Association 94*, 247–253.

Gimenez, O., B. Morgan, and S. Brooks (2008). Weak Identifiability in Models for Mark-Recapture-Recovery Data. *Modeling Demographic Processes in Marked Populations. Environmental and Ecological Statistics, Springer, New York*.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Haberman, S. (1979). *Analysis of Qualitative Data: Vol. 2. New Developments.* Academic Press, New York.

Hagenaars, J. (1993). *Loglinear Models With Latent Variables.* Sage (University Paper Series).

Hartley, H. (1958). Maximum likelihood estimation from incomplete data. *Biometrics 14*(2), 174–194.

Heady, P., S. Smith, and V. Avery (1994). 1991 Census Validation Survey: Coverage Report. *Office of Population Censuses and Surveys, London: HMSO*.

Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *American Statistician 46*, 261–269.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association 88*, 1047–1060.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

International Working Group for Disease Monitoring and Forecasting (1995a). Capture-recapture and multiple-record systems estimations I: history and theoretical development. *American Journal of Official Epidemiology 142*, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple-record systems estimations II: applications in human diseases. *American Journal of Official Epidemiology 142*, 1059–1068.

Isaki, C. and L. Schultz (1986). Dual system estimation using demographic analysis data. *Journal of Official Statistics 2*(2), 169–179.

Judson, D. (2000). The Statistical Administrative Records System: System Design, Successes, and Challenges. *Paper presented to the National Institute of Statistical Sciences (NISS) Data Quality Workshop, Nov 30 - Dec 1, 2000*.

Judson, D. (2006). Demographic Coverage Measurement: Cna Information Integration Theory Help? *Paper presented at the Joint Statistical Meetings, Seattle - Washington, USA, 7-10 August, 2006*.

Keohane, N. (2008). Local Counts: the future of the census. New Local Government Network Publication. NLGN, First Floor, New City Court, 20 St Thomas Street, London, SE1 9RS.

King, R. and S. Brooks (2001). On the Bayesian analysis of population size. *Biometrika 88*(2), 317–336.

Lazarsfeld, P. and N. Henry (1968). *Latent Structure Analysis.* Houghton, Mifflin.

Lesthaeghe, R. and D. van de Kaa (1986). Twee demografische transities? [Two demographic transitions?]. *Bevolking: groei en krimp [Population: Growth and Decline]*, 9–24.

Lincoln, F. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. *Circular of the Department of Agriculture 118*, 1–4.

Lindley, D. and A. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–41.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data.* Second Edition, John Wiley & Sons.

Madigan, D. and J. York (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika 84*(1), 19–31.

Martel, L. and É. Caron-Malenfant (2007). 2006 Census: Portrait of the Canadian Population in 2006: Findings. *Statistics Canada Catalogue*.

Martin, D. (2007). Editorial: Census present and future. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(2), 263–266.

McCutcheon, A. (1987). *Latent Class Analysis.* Sage (University Paper Series).

Meng, X. and D. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Mulry, M. and R. Griffiths (1996). Comparison of CensusPlus and Dual System Estimation in the 1995 Census Test. *Paper presented at the Joint Statistical Meetings, Chicago - Illinois, USA, 4-6 August, 1996*.

Nandram, B. and D. Zelterman (2007). Computational Bayesian inference for estimating the size of a finite population. *Computational Statistics and Data Analysis 51*(6), 2934–2945.

Office for National Statistics (2004). 2001 Census: Manchester and Westiminster Matching Studies Report. Methodology Directorate, Office for National Statistics. Crown Copyright.

Office for National Statistics (2007). Travel Trends - A report of the International Passenger Survey. Office for National Statistics. Crown Copyright.

Pearson, K. (1900). On the criterion that a given of deviation from de probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*(50), 157–72.

Petersen, C. (1896). The yearly migration migration of young plaice into Limfjord from the German Sea. *Reports of the Danish Biological Station to the Ministry of Fisheries 6*, 1–48.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Expanded edition). *Copenhagen: Denmarks Paedagogiske Institut*.

Redfern, P. (1989). Population registers: some administrative and statistical pros and cons. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 152*, 1–41.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, 577–591.

Royall, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika 57*(2), 377–387.

Salgueiro, M. (2002). *Distributions of Test Statistics for Edge Exclusion for Graphical Models.* Ph. D. thesis, University of Southampton.

Sanathanan, L. (1972a). Estimating the size of a multinomial population. *Annals of Mathematical Statistics 43*(1), 142–152.

Sanathanan, L. (1972b). Models and estimation methods in visual scanning experiments. *Technometrics 14*(4), 813–829.

Schnabel, Z. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly 45*(6), 348–352.

Schulte Nordholt, E., M. Hartgers, and R. Gircour (2004). The Dutch Virtual Census of 2001, Analysis and Methodology, Statistics Netherlands, Voorburg/Heerlen, July, 2004.

Scott, A. and D. Holt (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association 77*(380), 848–854.

Seber, G. (1982). *The Estimation of Animal Abundance.* Second Edition, London: Griffin.

Simpson, L., J. Hobcraft, and D. King (2003). *The 2001 One Number Census and Its Quality Assurance: A Review.* Local Government Association (England and Wales) - LGA Publications.

Skerry, P. (2001). Counting on the Census? *Society 39*(1), 3–10.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman & Hall/CRC.

Smith, P. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics 44*(4), 1177–89.

Smith, P. (1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika 78*(2), 399–407.

Sobel, M. (1995). *The Analysis of Contingency Tables.* Handbook of Statistical Modeling for the Social and Behavioral Sciences, pages 251-310. New York: Plenum Press.

Statistics Commission (2004). Census and Population Estimates and the 2001 Census in Westminster: Final Report. Statistics Commission - Report No.22. Crown Copyright.

Steele, F., J. Brown, and R. Chambers (2002). A controlled donor imputation system for a one-number census. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 165*(3), 495–522.

Stuart, E. and D. Judson (2003). An empirical evaluation of the use of administrative records to predict census day residency. *Proceedings of the American Statistical Association, Section on Government Statistics*.

Stuart, E. and A. Zaslavsky (2002). Using administrative records to predict census day residency. *Case Studies in Bayesian Statistics 6*, 335–349.

Thandrayen, J. and Y. Wang (2009). A latent variable regression model for capture-recapture data. *Computational Statistics and Data Analysis 53*(7), 2740–2746.

Treasury Select Committee (2002). *First Report on the 2001 Census in England and Wales.* House of Commons Treasury Committee Publication. Parliamentary Copyright, The Stationary Office (London).

United Nations Economic Commission for Europe (UNECE) (2007). *Register-based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics.* United Nations Publications: New York and Geneva.

Van Deusen, P. (2002). An EM algorithm for capture-recapture estimation. *Environmental and Ecological Statistics 9*(2), 151–165.

Venzon, D. and S. Moolgavkar (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics 37*(1), 87–94.

Wang, Y. and J. Thandrayen (2009). Mutliple-Record Systems Estimation using Latent Class Models. *Australian and New Zealand Journal of Statistics 51*(1), 101–111.

Webber, R. (1977). *The National Classification of Residential Neighbourhoods: An Introduction to the Classification of Wards and Parishes.* Planning Research Applications Group, Centre for Environmental Studies (Great Britain).

Werker, H. (1981). Results of the 1980 US Census Challenged. *Population and Development Review 7*, 155–167.

Wittes, J. and V. Sidel (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases 21*(5), 287–301.

Wolfgang, G. (1989). Using Administrative Lists to Supplement Coverage in Hard-to-Count Areas of the Post-Enumeration Survey for the 1988 Census of St Louis. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Wolter, K. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association 81*(394), 338–346.

Wolter, K. (1990). Capture-Recapture Estimation in the Presence of a Known Sex Ratio. *Biometrics 46*(1), 157–162.

Womersley, J. (1996). The public health uses of the Scottish Community Health Index (CHI). *Journal of Public Health 18*(4), 465–472.

Yule, G. (1900). On the Association of Attributes in Statistics. *Philosophical Transactions of the Royal Society of London: Series A 194*, 257–319.

Zaslavsky, A. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association 88*(423), 1092–1105.

Zaslavsky, A. and G. Wolfgang (1990). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Zaslavsky, A. and G. Wolfgang (1993). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data. *Journal of Business and Economic Statistics 11*, 279–288.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference 18*(2), 225–237.

Zwane, E. and P. van der Heijden (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling 5*(1), 39–52.