UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

School of Social Sciences

Understanding and dealing with unit nonresponse during and post survey data collection

by

Julia D'Arrigo

Thesis for the degree of Doctor of Philosophy

November 2011

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
SCHOOL OF SOCIAL SCIENCES

Doctor of Philosophy

UNDERSTANDING AND DEALING WITH UNIT NONRESPONSE DURING AND POST SURVEY DATA COLLECTION

By Julia D'Arrigo

Nonresponse in sample surveys is a longstanding concern among social researchers and survey methodologists. In addition to potential biases in point estimates, nonresponse can result in inflation of the variances of such estimates. This thesis focuses on understanding and dealing with unit nonresponse in sample surveys during and post data collection. In particular it looks at modelling the process leading to nonresponse using call record data; developing weighting adjustments for clustered nonresponse; and investigating variance estimation methods in the presence of nonresponse. During data collection, effective interviewer calling behaviours are critical in achieving contact and subsequent cooperation. Recent developments in the survey data collection process have led to the collection of so-called paradata, which greatly extend the basic information on interviewer calls. The first part of the thesis develops multilevel models based on a particular type of paradata, call record data and interviewer observations, to predict the likelihood of contact and cooperation conditioning on household and interviewer characteristics. The research is based on the UK 2001 Census Link Study dataset. The results have implications for survey practice and, among others, inform the design of effective interviewer calling strategies, including responsive survey designs. Post-survey estimation methods to adjust and account for nonresponse, such as weighting methods, include inverse probability weighting and generalized raking estimation. The second part of the thesis investigates alternative inverse probability weighted estimators for clustered nonresponse through a simulation study. Results from an empirical application using data from the Expenditure and Food Survey 2001 are presented. It also discusses three forms of generalized raking estimator in the presence of nonresponse. Weighting methods might result in increased variability in the weights and thereby lower the precision of the survey estimates. This thesis explores alternative forms of linearization and replication variance estimators for generalized raking estimators under nonresponse that allow for variation in the weights.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Declaration of Authorship

I, Julia D'Arrigo, declare that the thesis entitled 'Understanding and dealing with unit nonresponse during and post survey data collection' and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. 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;
- parts of this work have been published as:

D'Arrigo, J. and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. *Survey Methodology*, **36**, 2, 181-192.

Durrant, G.B., D'Arrigo, J. and Steele, F. (2011). Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences. *Journal of the Royal Statistical Society*, Series A, **174**, 4, 1029-1049. An earlier version of this paper is available at http://surveymethodology.eu/conferences/warsaw-2009/presentation/301/

Skinner, C. and D'Arrigo, J. (2011). Inverse Probability Weighting for Clustered Nonresponse. *Biometrika*, **98**, 4, 953-66.

Durrant, G.B., D'Arrigo, J. and Steele, F. (2012). Modelling Interviewer Call Record Data Using Multilevel Event History Analysis: Understanding the Process Leading to Cooperation or Refusal. Conditionally accepted by the *Journal of the Royal Statistical Society*, *Series A*.

Signed:

Date:

Acknowledgements

This doctoral thesis would have never been possible without the contribution of a number of people and organisations.

First and foremost I would like to express my greatest gratitude to my thesis supervisors Prof Chris Skinner and Dr Gabriele B. Durrant for their guidance, support, encouragement and helpful comments throughout the development of this doctoral thesis. I would also like to thank Dr James Brown, my PhD internal examiner, for his very helpful and constructive comments to an earlier version of this thesis. My gratitude is extended to Prof Fiona Steele from the Centre for Multilevel Modelling, University of Bristol, whose remarks and suggestions were always very useful.

The research has been partly funded by the Southampton Statistical Sciences Research Institute (S3RI) and the UK Economic and Social Research Council (ESRC), 'Hierarchical analysis of unit nonresponse in sample surveys', grant number: RES-062-23-0458 and 'The Use of Paradata in Cross-Sectional and Longitudinal Research', grant number: RES-062-23-2997. Their support is gratefully acknowledged. I am especially indebted to Prof Peter W. F. Smith, Director of S3RI, for his outstanding support and encouragement, very much needed, during the last year.

This work contains statistical data from the Office for National Statistics (ONS) which is Crown copyright and reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. I am grateful to the ONS for making the data available.

Many friends and colleagues from all around the world had positive influence on this thesis and I am grateful to them for helping me get through the difficult times, and for all the emotional support, camaraderie, entertainment, and caring they provided.

Last but not least I thank my parents, José Luis and Cristina, and my sisters, Florencia and Virginia, who have always been there for me. Most of all, I wish to thank my loving and supportive husband, Alejandro, and my two wonderful children, Sofia and Santiago. Through their love, patience, care and unwavering belief in me over the last few years, I have been able to complete this long PhD journey.

Chapter 1

Introduction

1.1 Nonresponse in sample surveys

Nonresponse in sample surveys is a longstanding concern among social researchers and survey methodologists. Nonresponse occurs when sampled members do not provide the requested information for one or more survey variables or are not contacted during the data collection process. For example, Hansen and Hurwitz (1946) pointed out the problem with large nonresponse rates on mail questionnaire surveys and proposed following-up the mail attempts by taking a subsample of nonrespondents with face-to-face interviews. Kish (1965) observed that differences in response rates across subgroups may introduce bias into survey estimates. In addition to potential biases in point estimates, nonresponse can result in inflation of the variances of point estimates due to reduced sample sizes.

Nonresponse bias usually receives much attention in the survey literature (Groves, 2006; Olson, 2006) as it is the main reason that survey agencies dedicate great efforts to reduce and adjust for nonresponse. Nonresponse bias occurs when respondents differ from the nonrespondents with respect to the characteristics to be investigated. Nonresponse bias is an important menace to the validity of all survey estimates.

In recent decades, problems caused by nonresponse have increasingly concerned survey practitioners as many surveys appear to show a decline in response rates. Curtin et al. (2005) presented falling response rates in several United States household surveys; de Leeuw and de Heer (2002) found that response rates have been declining over several years across different types of surveys in 16 developed countries; Tourangeau (2004) reviewed three recent developments in survey methodology within the context of decreasing response rates for all types of surveys: new methods of telephone sampling, new theories regarding causes and consequences of nonresponse, and new modes of data collection.

In general two types of nonresponse behaviour can be distinguished: unit and item nonresponse. Unit nonresponse occurs when eligible sample units fail to respond to a survey, e.g. because of noncontact, explicit refusal to cooperate or other reasons such as language barrier. Item nonresponse occurs when responding units do not answer some of the survey questions. The focus of this thesis is on unit nonresponse and does not further examine the concept of item nonresponse.

Unit nonresponse might be classified into three main components: noncontact, inability to respond and explicit refusal to cooperate. Noncontact includes both the failure of the interviewer to locate the sample unit and the failure to make contact with the sample unit. For example, noncontact may refer to the interviewer inability to talk to a responsible resident at the sampled household in a face-to-face or telephone survey. Those who fail to respond to a survey due to reasons such as ill health, infirmity and language barrier are classified as unable to respond. The last category refers to those who clearly refuse to participate in the survey after contact has been made.

There are two broad areas of research on dealing with survey unit nonresponse which involve: (a) strategies prior and during data collection to enhance response rates, including response survey designs and follow-up surveys; and (b) post-survey estimation methods that include some sort of adjustment to compensate for nonresponse.

Research to support (a) requires understanding nonresponse as a social phenomenon. This includes research looking at how nonresponse depends on individual and household characteristics as well as interviewer attributes. Goyder (1987), Groves and Couper (1998), Stoop (2005), and Durrant and Steele (2009) reviewed this vast literature and observed differences in response levels by characteristics such as age, geography and employment status of the household representative. Groves and Couper (1998) and Durrant and Steele (2009) also noted a quite distinct underlying nonresponse process for noncontact and refusal, observing that some predictors, such as employment status of the household representative, have opposite effects on the probability of noncontact and refusal.

Some efforts to increase response rates include incentives, more call attempts and follow-ups. Goyder (1987) showed that incentives are likely to result in higher response rates, even after controlling for survey design characteristics, such as length of the survey, sponsor or topic. Singer et al. (1999) found that monetary incentives were more effective in increasing response rates than gifts. They also observed that the effect of incentives is inversely proportional to the response rate: the lower the response rate

the larger the effect of an incentive. Goyder (1994), Hopkins and Gullickson (1992) and Singer et al. (1999) found that the impact of incentives given at the time of the survey request was greater than promised incentives. On the other hand, incentives can potentially introduce bias in the data by getting disproportionately more units from a select population subgroup into the respondent group (see Singer, 2002; Singer and Kulka, 2002 for general reviews of incentives).

Lynn et al. (2002) found that extended interviewer efforts, such as more call attempts and follow-ups, appear to reduce certain types of nonresponse biases due to increased contact rates. However, they also observed that greater interviewer efforts have limited effectiveness in reducing refusal rates and thus refusal bias. More recently, social researchers have been investigating the use of propensity models to predict the likelihood of response based on field process data (Kennickell, 2003; Sangster and Meekins, 2004; Groves and Heeringa, 2006; Bates et al., 2008). These data usually include call record information such as time and day of the call and outcome of the call. For face-to-face surveys, these data might also contain interviewer observations about the household and neighbourhood captured by the interviewer during data collection These models may inform the design of efficient and effective calling behaviours and follow-ups as well as responsive survey designs (Groves and Heeringa, 2006; Laflamme et al., 2008), where the continuous measurement and monitoring of the process and survey data offers the opportunity to alter the design during the course of the data collection to reduce costs and to increase the quality of the survey data. Propensity models might also be used to explore the role of the interviewers on survey nonresponse (Groves and Couper, 1998; O'Muircheartaigh and Campanelli, 1999; Pickery and Loosveldt, 2002; Durrant and Steele, 2009). Blom et al. (2010), for example, used a three-level logistic regression model to investigate the role that interviewers play in producing differences in response levels across countries in the European Social Survey.

Another factor affecting nonresponse in sample surveys during the data collection stage is the mode of data collection. Face-to-face surveys often have the highest response rates followed by telephone and mail surveys respectively. Web surveys have been rapidly embraced by the commercial research sector as a faster and cheaper mode of data collection despite serious concerns about coverage and nonresponse rates associated with these surveys (Couper, 2001). Tourangeau et al. (2000) examined the

complex psychological processes that make respondents more likely to cooperate to certain modes of data collection.

After data collection is completed, the second possibility of dealing with nonresponse in sample surveys is through post-survey modifications, such as weighting adjustments. Weighting methods are widely used to compensate for problems created by survey nonresponse. These methods are commonly used to compensate for unit nonresponse while imputation is typically but not exclusively employed to deal with item nonresponse. This thesis deals with weighting methods and does not consider imputation methods.

Weighting adjustments make use of auxiliary information to correct for nonresponse bias. The basic principle of weighting methods involves using an appropriate model based on auxiliary information to estimate response propensities for each unit in the sample. Then, these estimated propensities are used to adjust the probability-sample weights and produce estimates with lower biases. Estimation of response propensities assumes a stochastic approach to model survey nonresponse, which views the response units as the result of two probabilistic selections. First, a sample is selected from the finite population and then, the response units are realised as a subset of the sample. Further details about this stochastic approach can be found in Särndal and Lundström (2005). It is possible to distinguish between two types of auxiliary variables to use for weighting adjustment purposes are: (a) sample-based variables, i.e. variables known for the sampled units but not the entire population; (b) population-based variables, i.e. variables known for the entire population. Särndal et al. (1992) explored the use of these two types of variables via regression fitting.

There are different weighting adjustment procedures. One method, called inverse probability weighting, is to derive estimates of the response propensities from the sample units, and then to use the invert of these estimated probabilities as the weighting adjustments. As early as 1949, Politz and Simmons suggested a simple method to directly estimate contact probabilities. They proposed first to estimate the proportion of time each interviewed person was at home during the interviewing hours and divide questionnaires into six groups according to these estimates; and then, use the inverse of the group time-at-home estimate as the weighting adjustment factor. Bartholomew (1961) described another type of nonresponse adjustment to compensate for noncontact. He proposed to treat the second call successes as a random sample of all failures at the first call (other than those due to removals or deaths), and give

different weights to results from the first and second call. In 1986, Little observed that direct estimates of response propensities may result in unstable estimates if some of the estimated probabilities are close to zero. Little (1986) suggested sorting the sample by estimated response probabilities, forming five groups based on the quintiles of the response propensity distribution, and assigning the same weighting adjustment to all sampled units within a category. This weighting procedure is usually refers as weighting class adjustment. A more recent approach to estimate response propensities is by fitting parametric models, such as logistic or probit models, relating the study variable of interest and auxiliary variables (Cassel, Särndal and Wretman, 1983; Bethlehem, 1988; Fuller & An, 1998; Lundstrom and Särndal, 1999). Nonparametric methods, such as CHAID (Chi-square Automatic Interaction Detector; see Kass, 1980) and CART (Classification and Regression Trees; see Breiman, 1984), can also be used to estimate the response probabilities. Rizzo et al. (1996) compared the estimates obtained through several methods for adjusting weights, including nonresponse weight adjustments based on CHAID models, to estimates from independent sources. Da Silva and Opsomer (2004, 2006) investigated the properties of nonparametric methods that only require the response propensities to be related to the auxiliary variables by a smooth but unspecified function. Särndal and Lundström (2005) noticed the importance of powerful auxiliary variables to effectively model response probabilities and to reduce bias.

Other weighting adjustment procedure is calibration estimation. Calibration estimation guarantees that estimates based on data from a sample match previously determined benchmarks. The principle of calibration is to derive new weights by minimizing the total distance between the initial weights and the new weights, while ensuring that the new weights satisfy the benchmark requirements. Deville and Särndal (1992) introduced calibration estimation for the full response set-up and showed that estimators such as the generalised regression estimator and the poststratified estimator are special cases of calibration estimators. These calibration estimators can be modified and used to deal with unit nonresponse (Lundström and Särndal, 1999). Benchmarks for calibration may be obtained as estimates from a further sample, which may be considered sufficiently accurate, or from the population. Therefore, calibration estimators can be either sample-based or population-based.

Another special case of calibration estimator that is extensively used to adjust for survey nonresponse is the weighting class adjustment mentioned earlier (Little, 1986).

Weighting class adjustment consists on dividing the sample into a number of groups and giving each a weight equal to the inverse of its estimated response probability. The groups, usually referred to as cells, for the weighting class adjustment should be formed allowing for variables that are predictive of response and are correlated to the main statistics being produced. This approach can be limited when a large number of auxiliary variables are available. Two alternative calibration methods that allow including as many auxiliary variables as needed are raking (Deming and Stephan, 1940) and the two-way classification method (Särndal and Lundström, 2005). These methods only control to marginal totals.

Weighting adjustments can affect not only nonresponse bias but also the variance of an estimator. In fact, they can inflate the variance or they can reduce it (see, for example, Little and Vartivarian, 2005). It therefore becomes important to be able to estimate the variance of a weighted estimator in the presence of nonresponse. There are two methods commonly used to compute variance estimates for complex sample surveys: Taylor series linearization (see, for example, Wolter, 2007) and replication (see, for example, Fuller, 1998). These methods account for the nonresponse adjustments (see, for example, Valliant, 1993; Yung and Rao, 2000). If the variance estimate method does not account for the nonresponse adjustments, then the variance estimate might be underbiased resulting in short confidence intervals. Valliant (2004) studied through simulation the differences between linearization and replication methods to account for weighting adjustments on variance estimates. He reported that the linearization variance estimators were negatively biased and produced confidence intervals that cover at less than the nominal rate and that the jackknife replication estimator generally yields confidence intervals that cover at or above the nominal rate.

1.2 Purpose and outline of the thesis

This thesis centres on understanding and dealing with unit nonresponse in sample surveys during and post data collection. In particular it focuses on three specific themes: (1) modelling the process leading to nonresponse using call record data (Chapter 2); (2) developing weighting adjustments for clustered nonresponse (Chapter 3); and (3) investigating variance estimation methods in the presence of nonresponse (Chapter 4). The first objective relates to strategies that may be used prior and during

data collection to enhance response rates. The last two topics refer to post-survey estimation methods to adjust and account for nonresponse. All these approaches aim to correct for the potential biasing impact of nonresponse in point estimates and to minimise its effects on the associated variance estimates.

Chapter 2 deals with the two main types of unit nonresponse in sample surveys: noncontact and refusal. This chapter focuses on face-to-face surveys but some findings may also apply to telephone surveys. It first develops propensity models that predict the likelihood of contact in the field conditioning on household, interviewer and area influences. Then, it focuses on the process leading to cooperation and jointly models the different types of outcomes at each call using interviewer call record data and controlling for household and interviewer characteristics. The model allows for four different outcomes at each call: full or partial cooperation, refusal, making an appointment and other forms of postponement, such as appointment broken or the interviewer withdrew to try again later. These models investigate the usefulness of call record data and interviewer observations to predict the response outcome in six major UK face-to-face surveys.

Multilevel analysis (e.g. Steele et al., 2004) is used to model the probability of contact or cooperation at each call allowing for the hierarchical structure of the data with clustering of outcomes within household and clustering of households within a cross-classification of areas and interviewers. These models also account for unobserved household and interviewer characteristics. To model the process leading to contact a multilevel discrete time hazard model is used, conditioning on noncontact made prior to that call. To model the process leading to cooperation, conditioning on contact having been made with the household, a multilevel multinomial logistic regression analysis (e.g. Durrant and Steele, 2009) is employed. Multilevel models are motivated by a range of both technical and substantive reasons.

In Chapter 3, alternative inverse probability weighted estimators for clustered nonresponse are investigated. Cluster-specific non-ignorable (CSNI) nonresponse, as introduced by Yuan and Little (2007), is considered in this chapter. CSNI describes the case when nonresponse may depend on unobserved cluster random effects which may be correlated with the survey variables. Three standard forms of inverse probability weights are examined: response propensity weights (e.g. Little, 1988), weights based on predicted random effects (e.g. Durrant and Steele, 2009) and weights based on estimated fixed effects, where the random effects are treated as unknown parameters. A new

approach using conditional logistic regression is also proposed. The properties of the alternative estimators and associated variance estimators are investigated in this chapter through a simulation study and results from an empirical application are presented.

Chapter 4 reports a simulation study of the properties of alternative generalized raking estimators and associated variance estimators with respect to the effects of both sampling and nonresponse. The simulation study is designed to mimic two major European surveys: the UK Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). Three forms of generalized raking estimators are considered: the generalized regression estimator, the classical raking ratio estimator and the 'maximum likelihood' raking estimator (Brackstone and Rao, 1979; Fuller, 2002). The GREG estimator is widely used in many surveys, in particular in the context of nonresponse (Särndal and Lundström, 2005). The second estimator has been used in practice in the LFS and a version of the third estimator has been used in practice in the SIE. Alternative forms of linearization variance estimators (Demnati and Rao, 2004; Deville and Särndal, 1992), for generalized raking estimators are defined via different choices of the weights applied (a) to residuals and (b) to the estimated regression coefficients used in calculating the residuals. A grouped jackknife replication method, which recomputes weight adjustments for every replicate, is also examined to calculate an alternative variance estimator accounting for the nonresponse adjustments.

The final chapter of this thesis presents some concluding remarks.

Chapter 2

Modelling the process leading to nonresponse using call record data

2.1 Introduction

Establishing contact with eligible sample units is an essential part of the response process together with obtaining productive interviews. In recent years, these tasks have become progressively more difficult and so more expensive and time-consuming (Weeks et al., 1980; Groves and Couper, 1998; Cunningham et al., 2003). Increasing contact rates by scheduling calls when householders are more likely to be at home may not be productive if at these times refusals are more likely. Therefore, both mechanisms need to be understood to develop effective interviewer calling strategies that result in increased contact rates and subsequent higher cooperation rates. Even though survey agencies have become increasingly concerned with understanding and improving the data collection process, research so far has mainly investigated the final outcome, or specific call outcome, of contact/noncontact and cooperation/refusal rather than the process leading to these results. Weeks et al. (1980), for example, studied best time of day and day of the week to find someone at home in a 1976 US survey at the time of the first call. O'Muircheartaigh and Campanelli (1999) explored the influence of interviewers on refusals and noncontacts at the final outcome for wave 2 of the British Household Panel Survey. Durrant and Steele (2009) modelled the final survey outcome of refusal, noncontact or cooperation to investigate the effects of household characteristics on household unit nonresponse in six UK face-to-face government surveys. None of these studies allows the likelihood of contact or cooperation to vary across calls or examines how the call history may affect the outcome of future calls, which are some of the aims of this chapter.

To obtain information about survey data collection that might help to understand the response process, survey organisations have started to routinely collect call record data, such as day and time of the call, the outcome of the call and, in

particular for face-to-face surveys, observations made by the interviewers about the physical and social characteristics of the selected household and the neighbourhood. Such data are commonly referred as field process data or paradata (Couper, 1998), and greatly extend the basic information on interviewer calls. Paradata might also include additional information about the sample units from external records, such as presence of children or pensioners in the household. Paradata may be used in survey organisations to guide decisions on responsive or two-phase sampling designs (Groves and Heeringa, 2006; Eckman and O'Muircheartaigh, 2008), and also to obtain general knowledge about optimal calling practices to adequately schedule calls and follow-ups with the aim of increasing the probability of contact and cooperation (Purdon et al., 1999; Matsuo et al., 2006).

So far, analyses of paradata and interviewer calling strategies, in particular for face-to-face surveys, have been limited. For example, Weber and Burt (1972) and Weeks et al. (1980) examined best times of interviewer visits in face-to-face surveys. Greenberg and Stokes (1990) developed a set of rules for scheduling the time of the next call for a telephone survey conditioning on calling history. Kulka and Weeks (1988) investigated optimal calling protocols for telephone surveys based on the timing of previous calls. However, these studies examined average best times of day and days of the week to establish contact or cooperation without controlling for household or interviewer characteristics. These characteristics may have a significant impact on optimal interviewer calling strategies (Groves and Couper, 1998). Other studies controlled for basic information about the household or area, but without deriving household-specific estimates of the probability of contact or cooperation (Purdon et al., 1999; Groves and Couper, 1998; Brick et al. 1996; O'Muircheartaigh and Campanelli, 1999). Most research on best calling strategies has been carried out in the context of telephone surveys (e.g. Weeks et al., 1987; Greenberg and Stokes, 1990; Brick et al. 1996) rather than face-to-face surveys, although the latter offer a much wider range of observational information available for each household and call (Groves and Couper, 1998; Greenberg and Stokes, 1990). Previous empirical research which investigated the effect of a small number of factors influencing household unit nonresponse have often used simple methods such as descriptive analysis techniques or regression models that have ignored the hierarchical structure of the data where sample units are nested within interviewers (e.g. Purdon et al., 1999; Groves and Couper, 1996, 1998; Wood et al., 2006; Groves and Heeringa, 2006). Some studies have used multilevel modelling techniques to

analyse interviewer effects on various components of unit nonresponse; however, they were based on a single survey with a specific design and survey topic, a fairly small number of interviewers and households and a limited amount of information on household and interviewer characteristics (Pickery and Loosveldt, 2002, 2004; Pickery et al. 2001; O'Muircheartaigh and Campanelli, 1999). Hox and Leeuw (2002), used multilevel logistic regression analysis to examine the influence of interviewers' attitude on household survey nonresponse in different countries and several surveys; however, their models did not control for household characteristics that might be related to the likelihood of achieving cooperation.

This chapter illustrates the use of a particular type of paradata, interviewer call record and interviewer observation data, which are increasingly collected by survey organisations. It introduces the reader to the analysis of call record data in a multilevel modelling framework. The research presented in this chapter uses multilevel logistic analyses, which allows for clustering of households within interviewers, to separately study the process leading to contact and cooperation allowing for potential differences in the determinants of each type of nonresponse. There are technical and substantive advantages for using multilevel models over single-level models. Models that ignore the hierarchical structure of the data lead to underestimation of the standard errors of regression coefficients, in particular, of cluster-level variables, such as household and interviewer variables in this chapter (Snijders and Bosker, 1999; Goldstein, 2011). The standard error underestimation might lead to incorrect inferences about the effects of such variables. Among practical advantages, multilevel modelling allows exploration of substantive questions that is not possible in single-level models. For example, 'Is the extent of between-interviewer variation the same for contact and cooperation?' and 'Is the extent of between-household and between-interviewer variation the same for different types of call outcome?' This chapter aims to address some of these questions. The analyses use data from the Census Link Study, which provides an exceptional opportunity to analyse the effectiveness of interviewer calling behaviours and strategies to establish contact and obtain subsequent cooperation, controlling for household and interviewer characteristics. This study benefits from the availability of relatively rich paradata, including information recorded by the interviewer at each call to the household, interviewer observations about the household and neighbourhood, information about the interviewer-household interaction and detailed information about the interviewers themselves. The dataset combines call record data from six major UK

face-to-face surveys, which allow more general inferences to be made than in prior work. A key strength of these data is that individual and household characteristics from the UK 2001 Census are linked to the paradata for both respondent and nonrespondent households.

It is expected that this research will contribute to methodological progress in the analysis and modelling of call record data and the specification of suitable models to analyse such data. The findings may have important implications for survey practice, such as informing responsive survey designs, as defined by Groves and Heeringa (2006), effective interviewer calling behaviours, the design of call-backs and follow-ups of nonrespondents. Although survey organisations may not have access to information such as the control census variables considered in this study, the analysis provides useful information about the type of data that could be beneficial for predicting contact and cooperation and survey organisations could explore proxies for such variables from available data sources. It would also be possible to train interviewers to collect relevant observation data for each household and/or each visit to proxy such information.

This chapter is structured as follows. Section 2.2 describes the data upon which the research is based. Section 2.3 focuses on the process leading to contact and proposes a propensity model based on call record data and other paradata to predict the likelihood of contact at each call, conditioning on household and interviewer characteristics. The process leading to cooperation is studied in Section 2.4, modelling the response outcome at each call, conditional on contact having been made with the household at that call. A summary of the findings with implications for survey practice is provided in Section 2.5.

2.2 Data

2.2.1 UK 2001 Census Link Study

The research on this chapter is based on the UK 2001 Census Link Study dataset, which was produced by the UK Office for National Statistics (ONS), and includes the response outcome of six face-to-face major UK government surveys linked to household information from the UK 2001 Census, interviewer observations about the household, detailed information about the interviewers and area information from

aggregated census data. The dataset contains a total of 16,799 households (after excluding vacant and non-residential addresses, re-issues and unusable records, as described in Durrant and Steele, 2009), 565 interviewers and 392 areas defined at the local authority district level. The households included in the dataset were selected for interview in one of the six surveys during May-June 2001, the months immediately following the 2001 Census.

The six household surveys contained in the Census Link Study are the Expenditure and Food Survey (EFS), the Family Resources Survey (FRS), the General Household Survey (GHS), the Omnibus Survey (OMN), the National Travel Survey (NTS) and the Labour Force Survey (LFS). The surveys collect information based on the household as a whole and on the individuals within the households. The analyses in this chapter are based on household level data and individual level information was only used to derive variables recording information about the household reference person (HRP). The HRP variables facilitate moving from individual to household level, as every household has only one HRP. The EFS, created in 2001 by merging together the Family Expenditure and the National Food surveys, seeks to provide information on the pattern of spending and food consumption by households in the UK. The FRS, which has been carried out in Great Britain since 1992 and extended to include Northern Ireland in 2002, aims to provide information on living standards, people's relationship and interaction with the social security system. The GHS, created in 1971, is a multi-purpose survey which collects information from people living in private households in Great Britain on a range of core topics comprising, for example, family information, education, income, and demographic information about household members. The NTS, which has been running on an ad hoc basis since 1965 and continuously since 1988, aims to provide a comprehensive picture of personal travel behaviour. The OMN which began in 1990 is a multi-purpose survey which aims to obtain information about the general population or about particular groups. The questionnaire includes a set of core classificatory questions and a series of unrelated modules on varying topics at the request of customers. Core questions include information on demographic details, economic status, job details, employment status, full- or part-time working, and ethnic origin. The LFS created in 1979 aims to provide information about the UK labour market and unemployment. The survey seeks information on respondent's personal circumstances, their labour market status and income.

Figure 2.2.1: Refusal and noncontact rates for the six surveys in the Census Link Study dataset

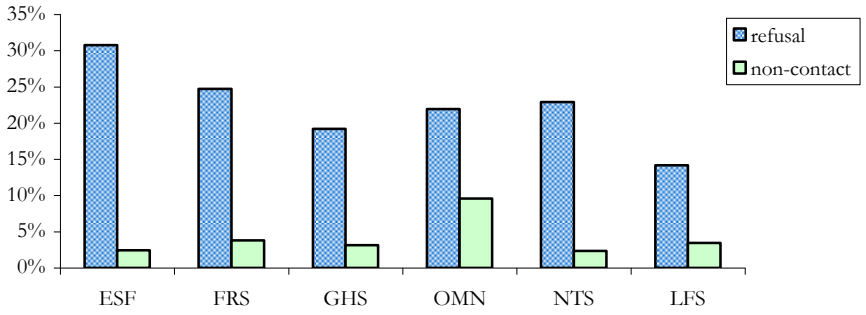
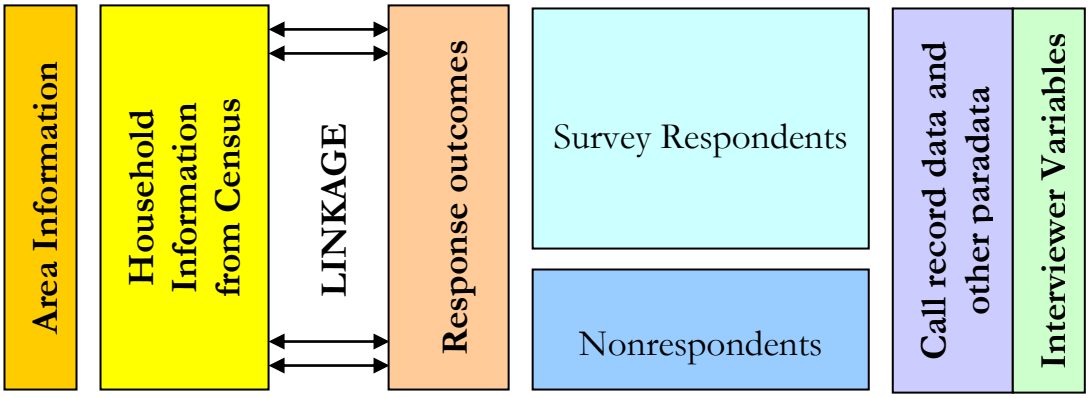


Figure 2.2.1 shows refusal and noncontact rates for each of the six surveys in the Census Link dataset. The noncontact rates for the surveys range from about 3% to 10%. A larger variation is observed among refusal rates across surveys, from around 15% for the LFS to 30% for the EFS, which may be explained by differences in the survey topic, interview length, length of data collection period, interviewer workload and additional requirements such as a diary. Although the noncontact rates might not appear very large in comparison to the refusal rates, establishing contact is a costly and time consuming process worthy of studying. Further details about these surveys can be found in Durrant and Steele (2009).

A great advantage of the Census link Study is that the survey data collected from the six surveys described above have been linked to the 2001 UK census records, available for both respondent and nonrespondent households chosen for interview (see Fig. 2.2.2). The 2001 census, which took place on 29 April, collected a varied set of information on the population, such as household accommodation, demographic characteristics (for example, gender, age, marital status), health and provision of care, pensioners households, dependent children, qualifications and employment, at that particular point in time. This data linkage provides an exceptional opportunity to investigate common characteristics of responding and nonresponding households. Another major benefit of this study is that relatively rich paradata at household level was also linked to the other data sources. These paradata is gathered by the interviewer during the data collection period of the six surveys in the study via an interviewer observation (IO) questionnaire (see Appendix A1). Further information about these data is presented on the following section. In addition, the Census Link Study includes detailed information about the interviewers and area information from aggregated census data, linked to the household level information. The area is defined as the local

authority district. The interviewer information was collected via a separate survey of all face-to-face interviewers employed for the UK Office for National Statistics (ONS) in 2001 (Interviewer Attitude Survey, IAS). The timing of the IAS was chosen to coincide with the UK Census in 2001 and was carried out prior to the surveys' fieldwork; however, some interviewers might have responded after the beginning of the fieldwork. Information on interviewers includes socio-demographic characteristics, and employment background, such as pay grade and experience, workload and planning, attitudes, strategies and behaviours for dealing with noncontacts and refusals as well as information about doorstep approaches.

Figure 2.2.2: The design of the Census Link Study 2001



The linkage of the various data sources with the response outcome of each survey, as illustrated in Figure 2.2.2, was carried out by the ONS. Linkage of the survey and census data was based on the address of the household, gender, age or date of birth and, if necessary, further identifying information. The linkage was carried out separately for every survey. About 95% of all households were successfully linked to their census record. The linkage of the interviewer observation data and the interviewer attitudinal data was based on the interviewer number. A number of quality checks and a significant amount of clerical review were carried out to identify and minimise any potential linkage errors. All linkage was quality assured by the ONS on the basis of comparisons of the distribution of key variables before and after the linkage. Possible effects of linkage errors have been described in Herzog et al (2007). Potential effects arising from both missing data and measurement error on multilevel models is discussed in Goldstein (2010). More detailed information about the rationale of the study, the data and the

linkage of the different datasets can be found in White et al. (2001), Durrant and Steele (2009) and Beerten and Freeth (2004).

2.2.2 Call record data and other paradata

This chapter focuses, in particular, in the usefulness of paradata to predict the likelihood of contact and cooperation. The available paradata in the Census Link Study contain records of calls, interviewer observations about the household and neighbourhood and information about the doorstep interviewer-household interaction. The call record data include the time and day of call, brief information on the contact strategy used at the call, and the outcome of the call. If contact with the household was achieved, the interviewer also captured information on age and gender of the main person talked to at each contact and whether this person made any comment or asked any question during the introductory conversation with the interviewer. The interviewer also recorded (usually at the first visit) their observations about the household and neighbourhood, such as type of accommodation, if there were any physical barriers to entry to the house, quality of housing and information about the household composition, such as any signs of the presence of children. The interviewer observation data are, in principle, available even if no contact was made with the household and might be used in practice as proxies for unknown census variables. Further call variables were derived for the analyses such as the time between calls, the number of noncontact calls (both prior to the first contact and in between two contact calls) and the number of previous contacts. Such variables, together with the call record variables, are call dependent (time varying) and so measured at the call level.

Some of the information captured by the interviewer via the interviewer observation questionnaire coincides with the information provided by the census (e.g. type of accommodation, indicator if children present). The models presented in this chapter use, wherever possible, the interviewer observation variables as these could always be observed and collected during the data gathering while access to census information is not usually possible. However, due to low quality of some interviewer observations (with large amount of missing data) compared to their census counterpart, some census variables, where available, might be included in the models.

The dataset contains 37879 calls made to establish first contact and a further 69619 calls after first contact was achieved, including intermediate noncontact calls

(noncontact calls after first contact was attained). The maximum number of contact calls made to one household is 13, which increases to 15 when noncontact calls are included. The median number of contact calls per household made by an interviewer (after first contact was established and excluding any intermediate noncontact calls) is 2 (and average is 2.5). The survey organisation provides calling protocol guidelines to each interviewer which indicates that the final response outcome for an address cannot be coded as 'noncontact' until at least four calls have been made. At least two of these calls should be in the evening or on a Saturday. Some general guidelines are also provided on how to avoid or deal with a refusal at the doorstep. The interviewer is strongly advised to call back at least once after a refusal.

2.3 Using paradata to predict best times of contact conditioning on household and interviewer influences

2.3.1 Introduction

This section focuses on the process leading to contact and aims to build response propensity models based on paradata to predict the likelihood of contact at each call, conditioning on household and interviewer characteristics. Discrete-time event history analysis (see, for example, Steele et al., 2004) is used to model the propensity of contact, allowing for household, interviewer and area effects in a cross-classified multilevel model. The model conditions on information available for each household, such as from administrative data and interviewer observations at prior calls, interviewer characteristics and call record data included as time-varying covariates. The key research questions are:

1. What are the best times of the day and days of the week to establish contact?
2. What are the best times to establish contact with certain types of households, in particular households that are generally more difficult to contact?
3. To what extent does establishing contact and the success of the timing of the call depend on interviewer characteristics?

2.3.2 Multilevel discrete time hazard model for the probability of contact

Multilevel event history analysis is used to model the probability of contact at a particular call, given that no contact was made prior to that call (i.e. model the number of calls to first contact). Households that are not contacted by the end of the data collection period have right-censored contact histories. The interviewer is said to have made contact with a household at a given call, the dependent variable in the model, if he/she was able to talk to at least one responsible resident at the sampled household, either face-to-face or through an entry phone.

Denote by $y_{i(jk)t}$ the binary indicator of contact, coded 1 if contact is made with household i by interviewer j in area k at call t and 0 if the contact attempt fails. The grouping of the j and k indices in parentheses, (jk), indicates a cross-classification of interviewers and areas, that is an interviewer may work in several areas and an area may be covered by several interviewers. The conditional probability of contact at call t given no contact before t – commonly referred to as the discrete-time hazard function – is defined as $\pi_{i(jk)t} = \Pr(y_{i(jk)t} = 1 \mid y_{i(jk)t-1} = 0)$. The multilevel cross-classified discrete-time hazard model, allowing for a clustering of households within a cross-classification of interviewers and areas, may be written

$$\log\left(\frac{\pi_{i(jk)t}}{1-\pi_{i(jk)t}}\right) = \alpha_t + \boldsymbol{\beta}' \mathbf{x}_{i(jk)t} + \boldsymbol{\delta}' \mathbf{z}_{i(jk)t} + u_j + v_k, \qquad (2.3.1)$$

where $\mathbf{x}_{i(jk)t}$ is a vector of time-varying covariates, with coefficients vector $\boldsymbol{\beta}$, including attributes of calls such as time and day of contact attempt t, number of calls made to the household prior to t, time of call at $t-1$ and two-way interactions between call and household-level variables. The vector $\mathbf{z}_{i(jk)t}$, with coefficient vector $\boldsymbol{\delta}$, includes time-invariant characteristics of households, from interviewer observations and the census; interviewers attributes and attitudes, from the Interviewer Attitude Survey; and area indicators, from aggregated census information. α_t is a function of the call number t ('time') which allows the probability of contact to vary across calls; here α_t is initially fitted as a step function, i.e. $\alpha_t = \alpha_1 D_1 + \alpha_2 D_2 + ... + \alpha_T D_T$ where $D_1, D_2, ..., D_T$ are dummy variables for calls $t = 1, ..., T$ with T the maximum number of calls, but simpler

monotonic functions are also explored. Unobserved interviewer and area characteristics are represented respectively by normally distributed random effects u_j and v_k : $u_j \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$.

After restructuring the data so that, for each household, there is a record for every contact attempt, the multilevel discrete-time event history model (2.3.1) can be estimated as a cross-classified model for the binary responses $y_{i(jk)t}$. Estimation is carried out using Markov chain Monte Carlo (MCMC) methods as implemented in the MLwiN software. MCMC methods are used in a Bayesian framework where every unknown parameter θ must have a prior distribution $p(\theta)$. The prior distribution quantifies the uncertainty in the values of the unknown model parameters before the data are observed. The default non-informative (also known as flat or diffuse) priors applied in MLwiN when MCMC estimation is used are: (1) for fixed parameters, $p(\beta) \propto 1$. This improper uniform prior is functionally equivalent to a proper Normal prior with variance c^2, where c is extremely large with respect to the scale of the parameter. An improper prior distribution is a function that is not a true probability distribution in that it does not integrate to 1; (2) for scalar variances, $p(1/\sigma^2) \sim \Gamma(\varepsilon, \varepsilon)$, where ε is very small. This proper prior is more or less equivalent to a Uniform prior for $\log(\sigma^2)$ (Browne, 2009; Rasbash et al., 2009). The Bayesian approach is used to effectively obtain maximum likelihood estimates of the unknown parameters. Other numerical approaches used later for estimating multilevel models, such as Gauss-Hermite quadrature or penalized quasi-likelihood (PQL), would not be feasible for the cross-classified model presented in this section. In addition, Rodriguez and Goldman (2001) found that quasi-likelihood approximate inference may result in a substantial underestimation of the fixed and random effects making this approach less attractive. In this section, results from 80000 chains with a burn-in of 5000 are presented; using approximate quasi-likelihood estimates (Goldstein, 2003) as starting values for the sampling.

To aid interpretation of the fitted model, predicted probabilities of contact are calculated for each value of the categorical covariates, holding constant the values of all other covariates in the model at their sample means. To obtain mean probabilities, this study averages across interviewer and area-specific unobservables by taking random draws from the interviewer and area random effect distributions. The simulation

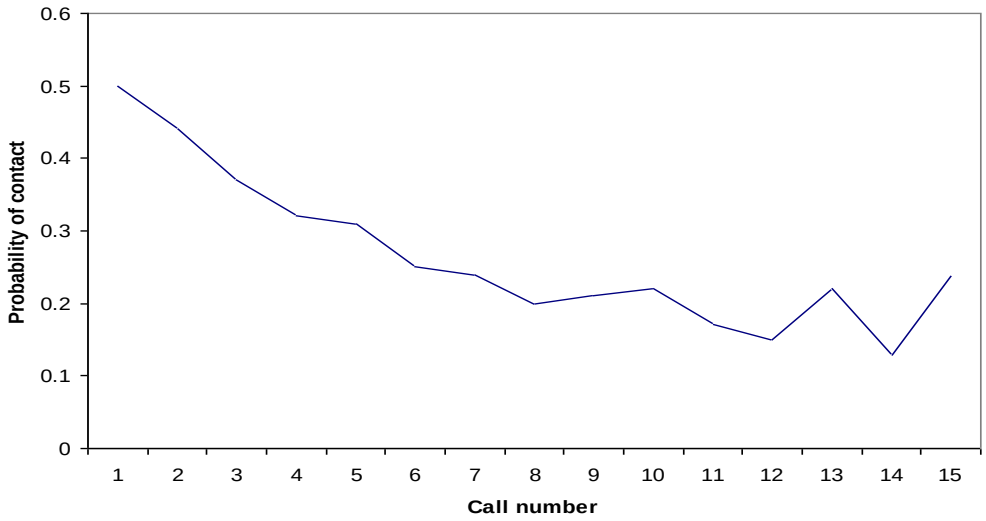
approach involves generating a large number of pairs of random effect values from independent normal distributions with variances $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$, calculating a predicted probability based on each pair of generated values and the estimated coefficients, and taking the mean across the simulated values. This procedure is implemented in MLwiN and described in Rasbash et al. (2009).

2.3.3 Results

The hazard rate and average best times of contact

This section presents some descriptive statistics on the contact process and results from preliminary models that informed the specification of the final multilevel model.

Figure 2.3.1: Estimated probabilities of contact for each call (hazard rate)†



† The sample sizes for calls 13-15 are less than 100 households.

Figure 2.3.1 shows the hazard of contact at each call, based on a simplified version of model (2.3.1) with only dummy variables for call number. In line with previous studies (e.g. Purdon et al., 1999; Groves and Couper, 1998), this figure shows a monotonic decline in the contact rate as the number of calls increases, until about call 9. The contact rate is highest at the first call, when about 50% of households were contacted, and decreases with each additional call. One possible explanation of this declining hazard might be attributed to the heterogeneity of contactability between

households. The slight increase in the contact rate for call 9 and 10, and the increase for calls 13 and 15, may also indicate that interviewers change their calling strategy and put in a greater effort to secure contact towards the end of their contact attempts. Another reason could be that interviewers have additional information that leads them to believe there is a chance of contact even after many failed attempts. It should be noted that from call 13 onwards the estimated probabilities of contact are based on fewer than 100 households. Based on the monotonic relationship between the probability of contact and call number, the specification of the baseline logit hazard, α_t in (2.3.1), is simplified by including the number of previous calls as a linear term.

Table 2.3.1: Probability of contact at first call, by day and time of call

		Contact probability	Total number of first calls made	% of all calls
Monday	Morning	0.46	682	4.1
	Afternoon	0.49	3310	19.8
	Evening	0.67	947	5.7
Tuesday	Morning	0.39	505	3.0
	Afternoon	0.48	2796	16.7
	Evening	0.63	810	4.8
Wednesday	Morning	0.36	327	2.0
	Afternoon	0.47	2176	13.0
	Evening	0.61	683	4.1
Thursday	Morning	0.44	290	1.7
	Afternoon	0.46	1864	11.1
	Evening	0.59	492	2.9
Friday	Morning	0.39	221	1.3
	Afternoon	0.42	1014	6.1
	Evening	0.57	286	1.7
Saturday	Morning	0.50	60	<1.0
	Afternoon	0.53	202	1.2
	Evening	0.43	51	<1.0
Sunday	Morning	0.50	10†	<1.0
	Afternoon	0.50	16†	<1.0
	Evening	0.67	9†	<1.0
Total		--	16799	100

Morning: 0.00-12.00, Afternoon: 12.00-17.00, Evening: 17.00-0.00
† indicates cells with a sample size of less than 30

Table 2.3.1 shows the probability of contact at the first call by time of day and day of the week. The most popular times to call are by far weekday afternoons, followed by weekday evenings and weekday mornings, with a clear decline in the frequency of calls from the beginning to the end of the week for all times of the day. It should be noted that due to interviewer working practices only few calls are made at the weekend, in particular on a Sunday. The highest contact probabilities can be found for evening calls, especially for Sunday to Wednesday evenings with a probability of more than 0.6. The chance of making contact in the evening decreases as the week progresses, with a comparatively low probability for Saturday evening of 0.43. On weekdays, the probability of making contact during the day is below 0.5, with a particularly low probability for Wednesday morning. For all weekdays, afternoons show a higher chance of contact than mornings. At the weekend the daytime contact probability is comparatively high at around 0.5.

The probability of contact at the second and third calls conditioning on the time of the previous call is also explored using descriptive statistics (Table 2.3.2 and 2.3.3 respectively). Due to small sample sizes, the time of day and day of the week variables were merged and categories collapsed to just four, i.e. weekend, weekday morning, weekday afternoon and weekday evening. The results may suggest that the best time for the second and third calls is a weekday evening, regardless of the time of the previous call, which supports earlier findings by Purdon et al. (1999), Groves and Couper (1998) and Kulka and Weeks (1988). The effect is greatest if the previous call was at a weekend and smallest if it was also made on a weekday evening.

Table 2.3.2: Probability of contact at second call conditional on timing of the previous call

Second call		First Call				
		Weekend	Weekday Morning	Weekday afternoon	Weekday Evening	Overall
Weekend	(239)	0.33	--	0.38	0.30	0.39
Weekday morning	(487)	--	0.31	0.35	0.28	0.34
Weekday afternoon	(3717)	0.35	0.34	0.37	0.39	0.37
Weekday evening	(3667)	0.65	0.54	0.53	0.50	0.53

-- indicates cells with a sample size of less than 30.
The number of second calls made per calling time are given in parentheses.
All cases where contact was made at the first call are excluded.

Table 2.3.3: Probability of contact at third call conditional on timing of the previous call

Third call		Second call				
		Weekend	Weekday morning	Weekday afternoon	Weekday Evening	Overall
Weekend	(213)	--	--	0.29	0.30	0.31
Weekday morning	(254)	--	--	0.27	0.25	0.28
Weekday afternoon	(1702)	0.36	0.34	0.30	0.23	0.28
Weekday evening	(2317)	0.59	0.47	0.47	0.39	0.44

-- indicates cells with a sample size of less than 30
The number of third calls made per calling time are given in parentheses.
All cases where contact was made at the first and second calls are excluded.

It should be noted that the ideal dataset for investigating best time of contact would be based on fully randomized calling times for all sample units. Such a design would, however, be impractical and very costly, at least for face-to-face surveys. It could be achieved to some extent for telephone surveys or experimental designs (Groves and Couper, 1998; Carrel and West, 2010). The dataset here, similar to previous work, provides information on observed calling times, i.e. the times that the interviewer chose to call on a household. If an interviewer's decision to call at a particular time can be regarded as independent of the characteristics of the sample unit, a departure from fully randomised calls should not be important. It seems reasonable to assume that interviewers choose when to make their first call with little, if any, prior knowledge about the sampling units. However, the timing of subsequent calls may depend on additional knowledge that the interviewer obtained at an earlier call. The models presented in this section attempt to adjust for this potential source of bias by controlling for information on the call history, interviewer observation variables and household information, which extends previous work on the analysis of call record data that did not include such controls (e.g. Bates et al., 2008). In practice, interviewer characteristics such as experience might also influence calling times, with more experienced interviewers more likely to be better at judging the most productive strategy for a given type of household. Thus, the models also attempt to control for differences between interviewer calling strategies by incorporating a number of interviewer characteristics in the model. The issue of non-random allocation of calling times to households has been discussed further in Purdon et al. (1999), Groves and Couper (1998) and Kulka and Weeks (1988).

The effect of day of the week and time of day is examined in a cross-classified multilevel discrete-time hazard model controlling for household, interviewer and area characteristics. The estimated coefficients for each category of the time of call variable

are provided in Table 2.3.4. The results confirm the indicative findings of Table 2.3.1, and largely support the conclusions of previous research, that evenings and weekends are optimal times to call (Weeks et al., 1980; Swires-Hennessy and Drake, 1992; Purdon et al. 1999; Groves and Couper, 1998). There is pervasive evidence that calling on weekday evenings yields the highest probability of contact, with a particularly high probability towards the beginning of the week and decreasing thereafter. Calling at the weekend, in particular on a Sunday, also leads to a higher probability of response, with Sunday evenings showing a similar pattern to early weekday (Mon-Wed) evenings. (Due to this finding and the very small number of calls made on a Sunday evening, this category was combined with 'early weekday evening' in later models, see Table 2.3.5). The next most successful times to call are weekday afternoons. Weekday mornings are generally the worst times to establish contact. During the week, afternoons are better than mornings but it is the other way round at the weekend.

Table 2.3.4: Estimated coefficients for the variable 'day and time of call' when included as a main effect only in the cross-classified multilevel discrete-time hazard model, controlling for household, area and interviewer characteristics, but without any interaction effects

		$\hat{\beta}$ (ste)
Monday	Morning	-0.861 (0.085)
	Afternoon	-0.756 (0.051)
	Evening	Reference
Tuesday	Morning	-1.084 (0.090)
	Afternoon	-0.800 (0.052)
	Evening	-0.063 (0.054)
Wednesday	Morning	-1.040 (0.101)
	Afternoon	-0.784 (0.055)
	Evening	-0.059 (0.055)
Thursday	Morning	-0.879 (0.102)
	Afternoon	-0.851 (0.058)
	Evening	-0.155 (0.059)
Friday	Morning	-0.998 (0.116)
	Afternoon	-0.871 (0.066)
	Evening	-0.187 (0.073)
Saturday	Morning	-0.419 (0.140)
	Afternoon	-0.682 (0.096)
	Evening	-0.508 (0.201)
Sunday	Morning	0.122 (0.527)
	Afternoon	-0.422 (0.336)
	Evening	0.645 (0.453)

Morning: 0.00-12.00, Afternoon: 12.00-17.00, Evening: 17.00-0.00

Results from Table 2.3.4 inform the categorisation of the calling time variable in the final model (Table 2.3.5) which distinguishes eight calling times: early week (Mon-Wed) and late week (Thu-Fri) morning, afternoon and evening and weekend daytime and evening.

Best times of contact for different types of households

So far the average best times to call on a household was considered. However, the chance of making contact at a given time of day will depend on the characteristics of the household. Groves and Couper (1998) provided a theoretical framework for understanding and studying household survey nonresponse. This framework identifies a number of important influences on the likelihood of contacting a sample household, including the timing and frequency of the calls, social environmental and socio-demographic attributes, at home patterns of the householders and the presence of physical impediments to gaining access to the household. Such attributes may be separated into factors that are under the control of the interviewer or survey organisation, such as timing of the call and interviewer contact strategies (e.g. leave a card), and factors outside their control, such as characteristics of the household or area (Purdon et al., 1999). This analysis aims to control for all of these effects. Previous studies that analysed overall best times to contact have found that evening and weekend calls are optimal (Weeks et al., 1980; Swires-Hennessy and Drake, 1992). A logical question to ask is which households have the highest chance of contact during the day, so that survey agencies may reserve evening and weekend times for more difficult cases. This research therefore investigates interactions between call times and household characteristics to determine best times of contact for particular households. Another interesting question for survey agencies is whether changing the timing of the call increases the likelihood of contact. Therefore, this research investigates the influence of the call history (see Purdon et al., 1999; Groves and Couper, 1998 and Kulka and Weeks, 1988). A separate indicator for the first call was included in the model and variables relating to earlier calls, such as the time of the previous call, were coded zero for the first call. This coding allows the coefficients of these call history variables to be interpreted as effects for second and subsequent calls.

This research now investigates the best times to establish contact with certain types of households, in particular those households that are generally more difficult to

contact. Table 2.3.5 presents parameter estimates of two multilevel discrete-time hazard models which take account of household and interviewer characteristics and interactions between time-varying variables and household and interviewer characteristics. Model A excludes census variables since these would not normally be available to a survey agency. Model B represents the final model which aims to understand the process leading to contact, including census information. Potential proxies for census variables from interviewer observation variables are also explored in this section. The inclusion of census variables reduces the DIC (Deviance Information Criterion, Spiegelhalter et al., 2002) by only a small amount (i.e. by 163 from 46936 for Model A to 46773 for Model B), indicating that a model based only on interviewer observation variables does not have much less predictive power than the full model. Furthermore, there are no differences in the direction of effects between the two models, implying that similar results can be obtained also in the absence of additional administrative data, i.e. when the survey agency can only rely on recordings by the interviewers to obtain information about nonresponding households.

From Table 2.3.5 it is observed that the probability of contact is highest for the first call. The highly significant negative coefficient for number of previous calls after the first call indicates a decrease in the odds of contact by 10% ([1-exp(-0.110)]*100=10%) for each additional call net of all other factors in the model, in line with the descriptive analysis shown in Figure 2.3.1. Non-proportional effects of covariates are tested by interacting each with number of previous calls, but there is no evidence to suggest that the effect of any variable differed across calls. In the following a distinction between interviewer observation and census variables is made, although in practice, at least some of the census variables could be substituted by variables based on interviewer observations. To aid interpretation of the interaction terms in the model and in an effort to illustrate how to maximise the likelihood of contact, predicted probabilities are provided in Table 2.3.6. (These have been calculated for call 1 but the pattern in probabilities is exactly the same for subsequent calls because the lack of interactions with the number of previous calls implies that all effects are constant across calls.)

Table 2.3.5: Estimated coefficients (and standard errors) for two multilevel cross-classified logistic models for contact: Model A without census variables and Model B with census variables

Variable (ref= Reference category)	**Categories**	**Model A** $\hat{\beta}$ $(ste(\hat{\beta}))$	**Model B** $\hat{\beta}$ $(ste(\hat{\beta}))$
Constant		0.011 (0.086)	-0.870 (0.111)***
Survey indicator (ref = EFS)	FRS	0.076 (0.054)	0.077 (0.050)
	GHS	0.052 (0.047)	0.022 (0.044)
	OMN	0.171 (0.049)***	0.064 (0.045)
	NTS	-0.026 (0.049)	-0.008 (0.046)
	LFS	0.682 (0.053)***	0.280 (0.057)***
Call Record Data (time-variant)			
Previous call indicator (ref= First call)	Call previously made	-0.645 (0.061)***	-0.550 (0.060)***
Number of calls previously made		-0.083 (0.009)***	-0.111 (0.009)***
Day and time of call (ref = Sun-Wed eve)	Mo-Wed am	-0.536 (0.144)***	-0.305 (0.196)
	Mo-Wed pm	-0.541 (0.084)***	-0.457 (0.115)***
	Thu-Fri am	-0.727 (0.208)***	-1.110 (0.284)***
	Thur-Fri pm	-0.792 (0.111)***	-0.625 (0.146)***
	Thu-Fri eve	-0.087 (0.113)	-0.118 (0.152)
	Sat-Sun am	-0.600 (0.379)	-0.282 (0.493)
	Sat-Sun pm	-0.281 (0.234)	-0.346 (0.306)
	Sat eve	0.053 (0.644)	-2.472 (1.651)
Time of previous call (ref= Weekday evening)	Weekend	0.704 (0.147)***	0.615 (0.141)***
	Weekday morning	-0.008 (0.104)	-0.018 (0.104)
	Weekday afternoon	0.175 (0.052)***	0.172 (0.052)***
Number of days between calls (ref= Same day)	1-3 days	0.095 (0.043)**	0.089 (0.042)**
	4-8 days	0.257 (0.046)***	0.245 (0.045)***
	9-14 days	0.332 (0.080)***	0.311 (0.080)***
	15+ days	0.428 (0.154)***	0.290 (0.155)*
Card/message left (ref= No card/message left)	Card/message left	0.104 (0.035)***	0.095 (0.035)***
Interviewer Observations (time-invariant)			
Security device (ref= security device visible)	No security device visible	0.210 (0.030)***	0.192 (0.031)***
Type of accommodation (ref= Not house, i.e. flat, mobile home, other)	House	0.467 (0.058)***	0.350 (0.057)***
Houses in area in good or bad state of repair (ref= Good)	Fair-Bad	-0.238 (0.052)***	-0.186 (0.050)***
House in a better or worse condition than others in area (ref= Better)	About the same	-0.127 (0.039)***	-0.068 (0.040)
	Worse	-0.308 (0.056)***	-0.272 (0.056)***
Dependent children present (ref= Not present)	Present	0.323 (0.059)***	----
Household-level variables from the Census (time-invariant)			
Age (household reference person) (ref= 16 - 34)	35 - 49	----	0.165 (0.033)***
	50 - 64	----	0.389 (0.038)***
	65 - 79	----	0.444 (0.069)***
	80 and older	----	0.535 (0.080)***
Household type (ref= Single household)	Couple household	----	0.425 (0.027)***
	Multiple household	----	0.402 (0.075)***
Pensioner in household (ref= No pensioner in household)	Pensioner in household	----	0.113 (0.082)
Person with a limiting long term illness present (LLTI) (ref= Not present)	Household with one or more people with LLTI	----	0.085 (0.055)
Dependent children present (ref= Not present)	Present	----	0.557 (0.054)***

Adults in employment (ref= No)	Yes	----	0.120 (0.064)**
Interviewer-level Variables (time-invariant)			
Pay grade (ref= Merit 1 and 2)	Interviewer and advanced interviewer	0.144 (0.038)***	0.079 (0.047)*
	Merit 3 and field manager	0.128 (0.043)***	0.129 (0.057)**
Interviewer qualification (ref= Degree or postgraduate, other higher education)	A levels	-0.110 (0.047)**	-0.148 (0.059)**
	GCSE, qualifications below this level, no qualification	-0.022 (0.035)	-0.032 (0.043)
Interviewer Age (ref= 50 years or more)	Under 50 years	-0.122 (0.056)**	-0.142 (0.062)**
Use phone to make appointment (ref= Always, frequently, sometimes)	Rarely, never	0.097 (0.033)***	0.103 (0.041)**
Interactions between interviewer observations and household characteristics			
Day and time of call * Dependent children present (ref=Sun-Wed eve and No dependent children)	Mo-Wed am * Children	-0.416 (0.131)***	-0.090 (0.126)
	Mo-Wed pm * Children	-0.256 (0.074)***	0.146 (0.069)**
	Thu-Fri am * Children	-0.260 (0.190)	-0.093 (0.187)
	Thu-Fri pm * Children	-0.191 (0.093)**	0.061 (0.090)
	Thu-Fri eve * Children	-0.043 (0.110)	-0.155 (0.098)
	Sat-Sun am * Children	0.187 (0.404)	-0.613 (0.358)*
	Sat-Sun pm * Children	-0.152 (0.230)	-0.116 (0.207)
	Sat eve * Children	0.063 (0.578)	-0.267 (0.524)
Day and time of call * Adults in employment (ref= Sun-Wed eve and No adults in employment)	Mo-Wed am * Yes	----	-0.552 (0.143)***
	Mo-Wed pm * Yes	----	-0.590 (0.080)***
	Thu-Fri am * Yes	----	-0.083 (0.202)
	Thu-Fri pm * Yes	----	-0.591 (0.103)***
	Thu-Fri eve * Yes	----	0.034 (0.118)
	Sat-Sun am * Yes	----	-0.381 (0.364)
	Sat-Sun pm * Yes	----	-0.028 (0.243)
	Sat eve * Yes	----	2.669 (1.518)*
Day and time of call * Household with a person with limiting long term illness (LLTI) (ref= Sun-Wed eve and No person with LLTI)	Mo-Wed am * LLTI	----	0.152 (0.118)
	Mo-Wed pm * LLTI	----	0.315 (0.069)***
	Thu-Fri am * LLTI	----	0.193 (0.166)
	Thu-Fri pm * LLTI	----	0.131 (0.087)
	Thu-Fri eve * LLTI	----	-0.045 (0.104)
	Sat-Sun am * LLTI	----	0.369 (0.297)
	Sat-Sun pm * LLTI	----	0.274 (0.199)
	Sat eve * LLTI	----	0.435 (0.536)
Day and time of call * Pensioner in household (ref= Sun-Wed eve and No pensioner)	Mo-Wed am * Pensioner	----	0.342 (0.153)**
	Mo-Wed pm * Pensioner	----	0.318 (0.088)***
	Thu-Fri am * Pensioner	----	0.629 (0.213)***
	Thu-Fri pm * Pensioner	----	0.246 (0.113)**
	Thu-Fri eve * Pensioner	----	0.034 (0.128)
	Sat-Sun am * Pensioner	----	-0.717 (0.385)***
	Sat-Sun pm * Pensioner	----	0.069 (0.265)
	Sat eve * Pensioner	----	1.600 (1.551)
Day and time of call * Indicator if house (ref= Sun-Wed eve and and Not house)	Mo-Wed am * House	-0.531 (0.139)***	-0.519 (0.145)***
	Mo-Wed pm * House	-0.258 (0.078)***	-0.191 (0.078)**
	Thu-Fri am * House	-0.338 (0.199)*	-0.158 (0.201)
	Thu-Fri pm * House	-0.035 (0.104)	0.065 (0.104)
	Thu-Fri eve * House	-0.040 (0.105)	0.048 (0.100)
	Sat-Sun am * House	0.106 (0.347)	0.311 (0.357)
	Sat-Sun pm * House	-0.065 (0.214)	-0.090 (0.214)
	Sat eve * House	-0.371 (0.567)	-0.110 (0.564)
Day and time of call * Indicator if house in a good or bad state of repair (ref= Sun-Wed eve and Good)	Mo-Wed am * Fair/Bad	0.012 (0.117)	0.036 (0.120)
	Mo-Wed pm * Fair/Bad	0.198 (0.066)***	0.150 (0.065)**
	Thu-Fri am * Fair/Bad	0.536 (0.163)***	0.631 (0.169)***
	Thu-Fri pm * Fair/Bad	0.243 (0.085)***	0.199 (0.085)
	Thu-Fri eve * Fair/Bad	0.157 (0.092)*	0.120 (0.090)
	Sat-Sun am * Fair/Bad	0.509 (0.327)	0.485 (0.327)

	Sat-Sun pm * Fair/Bad	-0.200 (0.202)	-0.144 (0.197)
	Sat eve * Fair/Bad	0.031 (0.496)	-0.168 (0.483)
Day and time of call * Time of previous call (ref= Sun-Wed eve and Weekday eve)	Mo-Wed am * Weekend	0.078 (0.408)	-0.007 (0.417)
	Mo-Wed pm * Weekend	-0.714 (0.223)***	-0.567 (0.224)**
	Thu-Fri am * Weekend	-0.552 (0.785)	-0.211 (0.766)
	Thu-Fri pm * Weekend	-0.189 (0.460)	0.003 (0.465)
	Thu-Fri eve * Weekend	-0.682 (0.459)	-0.675 (0.443)
	Sat-Sun am * Weekend	-0.240 (0.681)	0.065 (0.667)
	Sat-Sun pm * Weekend	-0.833 (0.306)***	-0.761 (0.297)**
	Sat eve * Weekend	-1.319 (0.587)**	-1.203 (0.580)**
	Mo-Wed am * Weekday am	0.090 (0.245)	0.098 (0.246)
	Mo-Wed pm * Weekday am	0.086 (0.135)	0.156 (0.137)
	Thu-Fri am * Weekday am	0.447 (0.298)	0.492 (0.301)
	Thu-Fri pm * Weekday am	-0.102 (0.168)	0.043 (0.170)
	Thu-Fri eve * Weekday am	0.379 (0.190)**	0.359 (0.185)**
	Sat-Sun am * Weekday am	0.574 (0.524)	0.438 (0.521)
	Sat-Sun pm * Weekday am	0.149 (0.521)	0.214 (0.508)
	Sat eve * Weekday am	0.014 (1.690)	-0.581 (1.628)
	Mo-Wed am * Weekday pm	0.163 (0.143)	0.211 (0.146)
	Mo-Wed pm * Weekday pm	-0.039 (0.067)	-0.009 (0.067)
	Thu-Fri am * Weekday pm	-0.063 (0.179)	-0.074 (0.183)
	Thu-Fri pm * Weekday pm	-0.034 (0.086)	0.014 (0.086)
	Thu-Fri eve * Weekday pm	0.025 (0.087)	-0.021 (0.083)
	Sat-Sun am * Weekday pm	0.772 (0.313)**	0.853 (0.313)***
	Sat-Sun pm * Weekday pm	-0.444 (0.205)**	-0.458 (0.201)**
	Sat eve * Weekday pm	0.108 (0.584)	-0.048 (0.607)
Interactions between interviewer observations and interviewer characteristics			
Day and time of call * Interviewer Age (ref= Sun-Wed eve and 50 years or more)	Mo-Wed am * under 50 yrs	0.096 (0.118)	0.108 (0.123)
	Mo-Wed pm * under 50 yrs	0.017 (0.066)	0.035 (0.067)
	Thu-Fri am * under 50 yrs	0.044 (0.171)	0.130 (0.171)
	Thu-Fri pm * under 50 yrs	-0.023 (0.087)	-0.012 (0.087)
	Thu-Fri eve * under 50 yrs	-0.194 (0.093)**	-0.204 (0.092)**
	Sat-Sun am * under 50 yrs	-0.776 (0.339)**	-0.716 (0.337)**
	Sat-Sun pm * under 50 yrs	0.061 (0.200)	0.029 (0.193)
	Sat eve * under 50 yrs	0.026 (0.443)	-0.142 (0.440)
Interviewer variance	--	0.089 (0.013)***	0.078 (0.011)***
Area variance	--	0.006 (0.005)	0.009 (0.005)*

The estimated coefficients and their standard errors are the means and standard deviations of parameter values across 80,000 Markov chain Monte Carlo samples, after the burn-in of 5000 and starting values from second order PQL estimation. The missing value categories have been suppressed to save space.

* significant at the 10% level
** significant at the 5% level
*** significant at the 1% level

Coding of time of call: am = 0.00-12.00, pm=12.00-17.00, eve= 17.00-0.00

Household and neighbourhood characteristics based on interviewer observations

Factors that are outside the direct control of the interviewer (Purdon et al., 1999), include characteristics of the household that indicate at home patterns of household member, socio-demographic characteristics and indicators of physical impediments to accessing the household. This study investigates the influence of variables that may be regarded as proxies for the time spent at home and lifestyle, such

as indicators of a single-person household, presence of dependent children and pensioners. Of particular interest is the effect of interviewer observation data as survey agencies should be able to collect this information for all households, including noncontacts. Such data are especially useful when no information from administrative data or census is available. Interviewer data (time-invariant) include information about physical barriers to accessing the household (e.g. a locked common entrance, locked gate or entry phone), the presence of security devices (e.g. security staff, CCTV cameras or burglar alarm), indications about boarded-up or uninhabitable buildings in the area, household composition, quality of the housing and how safe the interviewer would feel walking in the area after dark.

This research considers the effects of a range of interviewer observations. All of these variables are predictive of contact in initial modelling (i.e. before controlling for household and interviewer effects), which suggests such variables are useful for guiding the process of establishing contact in the field, in particular in the absence of additional administrative data, i.e. when the survey agency can only rely on recordings by the interviewers to obtain information about nonresponding households.

Table 2.3.5 shows the effects of variables that remained significant in the final model. As may be expected, houses with no security device visible - such as a security gate, burglar alarm, CCTV cameras or security staff - are easier to contact. An observation that can be relatively easily recorded by the interviewer is whether the household lives in a house or a flat. For almost all times, it is easier to establish contact with householders living in a house rather than a flat, and this is true even after controlling for household characteristics such as location, number of people in the household and presence of children. Interactions between interviewer observation variables and time of call are also explored, of which a number are found significant in initial modelling. Two interactions remain significant in the final model adjusting for all other household level characteristics; these are the interaction between timing of call and type of accommodation as well as state of repair of houses in the area. The interaction term between the timing of the call and the type of accommodation (Table 2.3.6) reveals that on afternoons, for any day of the week, it is easier to make contact with residents of houses than of flats. Householders living in flats are most likely to be contacted in the evenings and on Saturday and Sunday mornings. Contact is found on average to be more difficult when the interviewer recorded that houses in the area are in a fair or bad state of repair and that the house is in a worse condition than others in the

area (Table 2.3.6). The interaction term between timing of the call and state of repair of houses in the area provides some indication that the contact rate is better for houses in a fair or bad state of repair compared to houses in a good state of repair for Thursday-Sunday mornings. The fact that people living in fair or bad state of repair houses are more likely to be reach during late week morning might be due to these people being more likely to be unemployed or to be casual or shift workers and therefore at home more during the day. On the other hand, contact rate is better for houses in a good state of repair for Sun-Wed evening. These findings might indicate that state of repair of houses in the area may be regarded as a proxy for people in full-time employment (or people with children) more likely to be at home during the evening.

It is also found indication that contact is more difficult to establish if there are any boarded-up or uninhabitable buildings in the area or if the interviewer does not feel safe walking along in the area after dark However, none of these effects remain significant after controlling for other interviewer observations and household characteristics from the census. These variables could be indicators for social deprivation indices not significant once the model effectively controls for other household characteristics, such as type of household, employment, area.

It should be noted that interviewers are also asked to record indication of the presence of children, which is (at least in principle) the same information available from the census data. It was decided, however, to use the census variable in the final model due to the potential higher data quality and less item-nonresponse of this variable. (For an interpretation of the effect of this variable see the subsection '*Household characteristics from the Census*').

Two other call-specific variables that are under the control of the survey organisation, and that may determine best times of contact, are the timing of the previous call and the length of time since the last call. Considering the main effect of time of previous call only (without the interaction term with time of current call in the model) it is found that if the previous call is already a weekday evening call then establishing contact at the next call becomes increasingly less likely, indicating a potentially difficult to contact household. Some indications for a significant interaction term between time of current call and time of previous call are found (Tables 2.3.5 and 2.3.6). If the previous call is a weekend call, it seems advisable to call early during the week either in the morning or evening, or on a weekend morning. If the previous call is on a weekday afternoon, promising times to call are evening and weekend and Mon-

Wed mornings. If the previous call is made during the evening, calling again during the evening is the most likely to lead to contact, although in comparison to other previous calling times the contact rate for such repeated evening calls is smaller. It may be conclude that there is some indication for varying the timing of the call. Overall, however, evenings and weekends are reliably good times to call. This indicates that interviewers may have some (although limited) options in increasing contact rates by changing the time of the call, in particular if it is to an evening or weekend. Similar conclusions are drawn by Weeks and Kulka (1988), although they presented only descriptive statistics for the timing of the first three calls. Purdon et al. (1999) did not find a significant interaction between time of current and time of previous call. They concluded that if a household is repeatedly called upon during the evening the contact probability decreases, indicating a more difficult household. Groves and Couper (1998) did not find interpretable conditional effects of the timing of previous calls.

The effect of the number of days between calls (Table 2.3.5) suggests that leaving a few days between calls, ideally about one or two weeks, increases the probability of contact compared to returning on the same day. The increased probability of contact for call-backs after one or two weeks may reflect effects of additional knowledge about the household gathered by the interviewer at the earlier call which led them to adopt such a calling schedule. For example, interviewers may have found out from neighbours that the household was on holiday. Unfortunately, this type of information was not recorded for each call.

Household characteristics from the Census

It is well known that single-person households, households without children or with primarily young people, and households in urban areas and in flats are the most difficult to contact (Durrant and Steele, 2009; Groves and Couper, 1998), and the results presented here confirm these findings (see also Table 2.3.6).

From Table 2.3.6, it can be observed that for almost all call times the probability of contact is higher for households with children, with particularly high probabilities on weekday evenings, all afternoons and Mon-Wed mornings. The fact that weekday afternoons are good times may be related to children being back home from school. For households without children, calls made on weekdays during the day are the least likely to result in contact, whereas weekday evenings are the most promising. In practice, indications of the presence of children may be obtained via interviewer observations (as

in Model A) or, at least in some countries, from administrative or register data, such as from child benefit records (for an example see Cobben and Schouten, 2007). Although estimated coefficients for time of call and dependent children present (Table 2.3.5) somewhat differ between Model A and B, the predicted probabilities of contact obtained using Model A (not shown here), for the two-way interaction involving these two variables, display the same pattern as those in Table 2.3.6 based on Model B. Therefore, using either the Census variable or its proxy from the interviewer observation data leads to the same modelling conclusions.

As might be expected, the contact rate for weekday mornings (Mon-Wed) or afternoons (Mon-Fri) is higher for households without any adults in employment than for households with at least one employed resident (Table 2.3.6). The reverse effect is found for the evenings. For households with adults in employment the probability of contact for both weekday and weekend evenings are higher than for households in unemployment. There is a lower chance of contact for households with adults in employment on weekend mornings than for households in unemployment but weekend afternoons perform very similarly. The contact rate for Saturday evenings is higher for households with employed adults than for those with no one in employment. (An indicator of whether any adults are in employment is also available from the interviewer observation questionnaire and could be used as proxy for the census variable. Again due to the higher data quality of census data the census measure is included in the final model. For an example where information on employment status and unemployment benefits is available from administrative sources see Cobben and Schouten, 2007.)

Table 2.3.6: Predicted probabilities† of contact (in %) for two-way interactions

Interaction between day and time of call and type of accommodation			
		Type of accommodation	
		House	Flats, other
Day and time of call	Mon, Tue, Wed morning	38.2	42.2
	Mon, Tue, Wed afternoon	42.3	38.6
	Sun, Mon, Tue, Wed evening	58.1	49.6
	Thu, Fri morning	28.6	24.9
	Thu, Fri afternoon	44.5	34.8
	Thu, Fri evening	56.4	46.7
	Sat, Sun morning	58.8	42.7
	Sat, Sun afternoon	47.5	41.2
	Sat evening	9.9	7.9

Interaction between day and time of call and state of repair of houses in area			
		State of repair of houses in area	
		Good	Fair-Bad
Day and time of call	Mon, Tue, Wed morning	47.9	44.3
	Mon, Tue, Wed afternoon	44.2	43.4
	Sun, Mon, Tue, Wed evening	55.4	50.8
	Thu, Fri morning	29.5	39.3
	Thu, Fri afternoon	40.2	40.5
	Thu, Fri evening	52.5	50.9
	Sat, Sun morning	48.5	55.8
	Sat, Sun afternoon	46.9	39.0
	Sat evening	9.8	7.1

Interaction between day and time of call and dependent children in household			
		Dependent children present	
		Present	Not present
Day and time of call	Mon, Tue, Wed morning	54.6	43.2
	Mon, Tue, Wed afternoon	56.6	39.6
	Sun, Mon, Tue, Wed evening	63.9	50.6
	Thu, Fri morning	35.3	25.7
	Thu, Fri afternoon	50.5	35.7
	Thu, Fri evening	57.5	47.7
	Sat, Sun morning	42.4	43.8
	Sat, Sun afternoon	53.0	42.2
	Sat evening	10.7	8.2

Interaction between day and time of call and time of previous call					
		Time of previous call			
		Week end	Wkday am	Wkday pm	Wkday eve
Day and time of call	Mon, Tues, Wed morning	60.7	48.0	55.4	46.1
	Mon, Tues, Wed afternoon	43.5	45.7	46.3	42.4
	Sun, Mon, Tues, Wed evening	67.7	53.0	57.6	53.5
	Thu, Fri morning	36.6	38.2	29.9	27.9
	Thu, Fri afternoon	53.3	39.0	42.8	38.4
	Thu, Fri evening	49.1	58.8	54.3	50.6
	Sat, Sun morning	62.9	56.8	70.4	46.6
	Sat, Sun afternoon	41.5	49.8	38.3	45.1
	Sat evening	5.3	5.3	10.3	9.2

Interaction between day and time of call and adults in employment			
		Adults in employment	
		No adult	1+ adult
Day and time of call	Mon, Tue, Wed morning	50.8	40.4
	Mon, Tue, Wed afternoon	47.1	36.0
	Sun, Mon, Tue, Wed evening	58.6	61.0
	Thu, Fri morning	31.9	32.7
	Thu, Fri afternoon	43.0	32.2
	Thu, Fri evening	55.3	59.0
	Sat, Sun morning	51.4	45.0
	Sat, Sun afternoon	49.8	52.0
	Sat evening	10.9	65.5

Interaction between day and time of call and pensioner in household			
		Pensioner in household	
		Present	Not present
Day and time of call	Mon, Tue, Wed morning	56.3	45.2
	Mon, Tue, Wed afternoon	52.0	41.5
	Sun, Mon, Tue, Wed evening	55.4	52.6
	Thu, Fri morning	43.7	27.2
	Thu, Fri afternoon	46.1	37.6
	Thu, Fri evening	53.3	49.7
	Sat, Sun morning	31.8	45.7
	Sat, Sun afternoon	48.6	44.2
	Sat evening	34.6	8.9

Interaction between day and time of call and person with limiting long term illness (LLTI)			
		Person with LLTI	
		Present	Not present
Day and time of call	Mon, Tue, Wed morning	51.4	45.6
	Mon, Tue, Wed afternoon	51.7	42.0
	Sun, Mon, Tue, Wed evening	55.1	53.1
	Thu, Fri morning	39.9	27.6
	Thu, Fri afternoon	43.1	38.0
	Thu, Fri evening	51.1	50.2
	Sat, Sun morning	57.2	46.2
	Sat, Sun afternoon	53.4	44.6
	Sat evening	14.2	9.0

Interaction between day and time of call and interviewer age			
		Interviewer age	
		Under 50 years	50 years or more
Day and time of call	Mon, Tue, Wed morning	49.6	50.4
	Mon, Tue, Wed afternoon	44.1	46.7
	Sun, Mon, Tue, Wed evening	54.4	57.8
	Thu, Fri morning	31.4	31.6
	Thu, Fri afternoon	39.0	42.6
	Thu, Fri evening	46.5	55.0
	Sat, Sun morning	31.0	51.0
	Sat, Sun afternoon	46.7	49.4
	Sat evening	8.3	10.8

† Predicted probabilities are calculated by varying the values of the two interacting variables, holding all other covariates at their sample mean value. In the case of a categorical variable, the dummy variable associated with a particular category takes on the value of the sample proportion in that category instead of the usual 0 or 1 value.

The call indicator variable has been fixed for call 1 to obtain these predicted probabilities but the trend in predicted probabilities would be the same for subsequent calls since interactions with the call-variable were not included.

Coding of time of call: morning (am)=0.00-12.00, afternoon (pm)=12.00-17.00, evening (eve)=17.00-0.00

The interviewer also has a good chance of finding someone at home during the week if there is at least one pensioner present. Particularly high probabilities of contact are observed during the day in the early part of the week. Weekday evenings are also good times to establish contact with pensioners. Compared to other types of households, the contact rate for households with pensioners is relatively low at the weekend, particularly mornings. This may be partially explained by older people being

more likely to have religious or family commitments on a Sunday for example. For households without a pensioner weekday evenings and weekend mornings are the best times to call. There is also a suggestion of a similar effect for households with an older household representative, where householders older than 50 are more easily contacted during the day on weekdays whereas the daytime contact rate is quite low for householders younger than 35; however, this effect is not significant any more once we controlled for the interaction effect of pensioners. For any time and day, it is found that the older the household representative the more likely it is to establish contact, whereas householders aged below 35 are the most difficult to contact (Table 2.3.5).

From Table 2.3.5 it is observed that the number of people in the household has a significant effect on contact, with larger households being easier to contact than single-person households. This may be expected since it will be more likely to find at least one person at home for larger households. The interaction between timing of call and number of people in the household is significant in initial modelling, but not after controlling for other markers for at-home patterns such as the presence of children and household members in full-time employment.

Households with at least one person with a limiting long term illness (LLTI) have high probabilities of contact throughout the week as would be expected since such persons may be more likely to be at home due to their restricted daily activities and some may have a carer present. The probability of contacting these households is particularly high during the week (Mon-Wed), which is almost as good a time to call as evenings and weekends. In preliminary analysis, a very similar effect for households with a carer present is found, but due to collinearity with the LLTI variable this variable is not included in the final model. Information on the presence of carers or persons with a long-term illness may be available in register or administrative databases (for an example see Cobben and Schouten, 2007). Alternatively, some crude proxies or indicators may be captured by the interviewer, for example via observations regarding wheelchair access to the house or a disabled parking permit visible in the car.

Geographical location and type of area are usually regarded as important predictors of noncontact (Groves and Couper, 1998). However, after controlling for household characteristics, such as household type, the London and urban-rural indicators are no longer significant. Interactions between the geographical variables and the timing of the call are also explored. The interaction with the London indicator is significant in a simple model, but not after adjusting for all household effects and their

interactions. In the absence of household-level information knowledge about geographical location and type of area (urban-rural), which can be easily observed and collected by the interviewer, may be regarded as proxies for such household information and are expected to be predictive of contact. In addition, area-level variables, such as long-term unemployment rate, percentage of older people and children and percentage of houses are all found significant in predicting noncontact before controlling for household and call-level information, but not in the final model. This implies that area variables may be also considered as proxies for household characteristics, in line with the findings of O'Muircheartaigh and Campanelli (1999).

The above findings are based on a pooled analysis of six UK surveys which are expected to differ in their contact rates, for example because of differences in their design, such as length of data collection period. It is found that the LFS has a significantly higher probability of contact than the other surveys considered. This may be due to a number of factors, such as LFS interviewers working only on that survey whereas normally interviewers may be expected to work on several surveys. They also have a comparatively lower workload, in terms of the number of addresses, and receive more intensive interviewer training, although it should be noted that the LFS also has shorter data collection period than the other surveys.

Influences of the interviewer on the process of contact

There is significant, although small, variation between interviewers in their contact rates in all models. Inclusion of the interviewer characteristics reduces the between-interviewer variance from 0.11 to 0.08, explaining about 27% of the interviewer variance. The relatively small between-interviewer variance indicates that even though interviewers play a significant role in the process leading to contact, the effect of their unknown characteristics might not be strong on the log-odds of contact. The between-area variance is substantially smaller than the between-interviewer variance and, controlling for household-level and call-level variables halved the between-area variance; in the final model area effects are only marginally significant at the 10% level (see Table 2.3.5).

Before looking into the interviewer characteristics that influence the contact process, it is important to note the potential problem of interpreting interviewer effects that may be confounded with area effects. In clustered survey designs an interviewer is normally assigned to one primary sampling unit (PSU) and their workload consists of all

sampled households in that PSU. To account at some degree for this potential confounding of area and interviewer effect, it is possible to employ an interpenetrated sampling design, with interviewers allocated at random to households (O'Muircheartaigh and Campanelli, 1999; Schnell and Kreuter, 2001). However, due to the high costs involved with implementing an interpenetrating design, this approach is rarely used in practice. Some previous studies with no such design ignored area effects in the research or area information was not available at all (e.g. Pickery and Loosveldt, 2004). The six surveys included in this study did not employ randomised interpenetrated sampling designs; however, a complete confounding of area and interviewer effects was avoided because most interviewers work on a number of surveys and some interviewers work across PSUs. In addition, the model allows for random area effects where areas are defined as local authority district level, a geographical area considerably larger than a PSU. As a result, interviewers and areas are cross-classified, i.e. an interviewer may work in several areas and an area may be covered by several interviewers. For other examples of the use of multilevel cross-classified models and a detailed discussion of different forms of (partial) interpenetrated sampling designs see Durrant et al. (2010) and von Sanden (2004), respectively.

Purdon et al. (1999) and Groves and Couper (1998) considered the role of the interviewer in establishing contact. They argued that after adjusting for the timing of the call the interviewer should not play a significant role. Groves and Couper (1998) nevertheless investigated if there are any further net effects of interviewer characteristics and explored simple relationships between interviewer attributes and the probability of contact. This study investigates the effects of a number of interviewer characteristics in an attempt to explain the between-interviewer variance in contact rates, including socio-demographic characteristics, experience and work background and interviewer strategies. It may be argued that more experienced and higher qualified interviewers may be better at establishing contact (for a preliminary analysis see Groves and Couper, 1998, p. 95). This research finds pay grade of interviewers to be an important factor in explaining part of the differences between interviewers, with interviewers in higher pay grades being better at establishing contact. A similar effect was found in Purdon et al. (1999) - although contrary to their a priori hypothesis of no interviewer effects after controlling for the timing of the call. It is also found that interviewers with a higher qualification such as a University degree or postgraduate education have higher contact rates. This may indicate that certain types of interviewers may be better at judging the best times to

call, for example through gathering information about the household from observation and talking to neighbours, and using such information to tailor their calling strategy to maximise the chance of contact.

The model also shows that older interviewers (50 years and over) are more successful at establishing contact which may possibly reflect their greater experience or the fact that they may appear more trustworthy. Another explanation may be that older interviewers may have fewer time-constraining commitments outside their job, such as looking after young children, allowing greater flexibility on calling times. The interaction between age of the interviewer and timing of the call (see Table 2.3.6) is also explored, and some evidence is found that older interviewers may be better in judging the best timing of the call for certain types of households: older interviewers are more likely than younger interviewers to achieve contact on weekday evenings, in particular Thursday and Friday, and on weekend mornings.

Slightly surprisingly, it is not found any significant main or interaction effects of the number of years of interviewer experience after controlling for the timing of the call as well as household and area characteristics, even if this is the only interviewer level effect in the model. This is in line with Groves and Couper (1998) who also did not find an effect of interviewer experience. The expected positive association between experience and the probability of contact might be more adequately captured by pay grade and qualification and, to some extent, age which are all found to be significant. It may be argued that the pay grade of the interviewer captures a combination of length of experience and interviewer performance, with better performing interviewers expected to be on higher pay grades. This combination of characteristics may therefore be more important in explaining differences between interviewers rather than simply the length of time an interviewer has been in the job (for a similar effect on refusal see Durrant et al., 2009).

Since survey agencies are particularly interested in behavioural differences between interviewers, it is also explored to what extent interviewer strategies influence the probability of contact. It is found that interviewers who report that they at least sometimes wait to explain the survey, rather than simply leaving behind information, are more likely to establish contact (Table 2.3.5), which suggests that interviewers who put in more effort and dedicate more time to each sample unit may be more successful at securing contact. It is also found that interviewers who always or frequently use the phone to establish contact, rather than visiting the household in person, perform worse

than interviewers who rarely or never use the phone. Again, this variable may be an indicator of interviewer effort. Somewhat surprisingly some interviewer strategies, such as how often they check with neighbours, are not found to explain differences amongst interviewers. However, it should be noted that these measures of interviewer practice are self-reported rather than from direct observation. This non-significant effect may have been caused by the fact that most interviewers responded to these types of questions in a similar way. This may highlight a potential downside of self-recorded interviewer behaviour. As suggested by Groves and Couper (1998), in the context of interviewer effects on cooperation given contact, it may be preferable to ask interviewers to record their strategy for each call or household. Some support for their recommendation is found: the variable indicating whether it is the interviewer's general practice to leave a card or message behind has no significant effect on contact, while the time-varying covariate capturing the same information for each call is found to be significant, showing an increase in the probability of contact at the next call if a card or message was left (see Table 2.3.5).

It may be argued that more experienced interviewers and interviewers on higher pay grades are better at establishing contact with harder-to-reach households. Effects of this type could help to inform the allocation of certain interviewers to potentially more difficult households. Therefore interaction effects between interviewer characteristics and type of household are explored, focusing on households that previous research identified as being harder to contact, such as single households, younger people or households without children. However, none of the effects explored are found to be significant after controlling for the timing of the call and household characteristics. Also a number of other interviewer characteristics considered are not found to be associated with the probability of contact, including gender of the interviewer, whether they worked for another survey organisation or had other paid employment, and indicators of whether the interviewer is happy to travel, to work evenings and weekends, or to stay overnight.

2.4 Modelling the process leading to cooperation or refusal

2.4.1 Introduction

This section builds on the research presented above by focussing on the next step in the response process: cooperation and refusal. The research presented here aims to analyse the process leading to cooperation or refusal. It jointly models the different types of outcomes at each call conditional on contact being made with the household by using multilevel multinomial logistic regression analysis (see, for example, Pickery and Loosveldt, 2002). The models control for household characteristics and also allow for the influence of the interviewer on the cooperation stage. The key research questions are:

1. What is the process leading to cooperation/refusal? Do call time-variant variables influence this process?
2. Are interviewer observations useful to predict cooperation?
3. To what extent does cooperation depend on doorstep interviewer-householder interactions?
4. What are best times to establish cooperation? Are these times affected by the outcome of previous calls?

2.4.2 Multilevel multinomial logistic model for the response outcome

Multilevel multinomial logistic analysis is used to model the response outcome at call t, conditional on contact having been made with the household at that call. The dependent variable in this study is defined as an indicator of other possible outcome versus cooperation, conditioning on contact made with the household. Household cooperation is defined as an interview carried out by at least one member of the household. (This study does not distinguish between full cooperation, where the whole household responds, and partial cooperation where only some household members respond). Other possible outcomes at each call are divided into three main components of nonresponse and defined as:

(1) refusal, household refused to participate in the survey

(2) appointment made, household made an appointment for the interviewer to come back at a different time/day

(3) other form of postponement, contact made with sampled household but not with a responsible resident, broken appointment, interviewer withdrew to try again later, e.g. he/she felt threatened.

The research is interested in the response outcome across all contact calls, not just until the first contact or the first time a form of cooperation is established with the household. However, some considerations are given to the process leading to first cooperation, i.e. the first instance of a cooperation outcome.

A multilevel model is used to account for the hierarchical structure of the data allowing for clustering of outcomes by household or interviewer due to unobserved household and interviewer characteristics. The nature of the data, one record per each contact call made to a household, makes it possible for an outcome to occur more than once within a household. For example, during the data collection period a householder might make several appointments, an appointment might be broken more than once, or different household members might refuse or cooperate to the interviewer at different calls. A multilevel model with household random effects allows for this possibility that the events of interest occur more than once to a household. Due to the complexity of the modelling, the large number of available household characteristics, and the findings in section 2.3.3 (page 37) that area characteristics are negligible once interviewer and household effects are controlled for, area level effects are not additionally included in the multilevel model presented in this section.

Denote by y_{tij} the outcome of call t ($t = 1,\ldots,T_i$) made to household i ($i = 1,\ldots,n_j$) by interviewer j ($j = 1,\ldots,J$) conditional on contact being achieved at t. The outcome of each call is coded as follows: 1 for a refusal, 2 if an appointment is made, 3 for other forms of postponement, and 4 for full or partial cooperation. The conditional probability of outcome s at call t given contact being achieved at t is denoted by $\pi_{tij}^{(s)} = \Pr(y_{tij} = s)$, $(s = 1,2,3)$. A multilevel multinomial logit model for the log-odds of outcome s $(s = 1,2,3)$ relative to outcome 4 (cooperation) may be written

$$\log\left(\frac{\pi_{tij}^{(s)}}{\pi_{tij}^{(4)}}\right) = \boldsymbol{\beta}^{(s)\prime}\mathbf{x}_{tij}^{(s)} + \boldsymbol{\delta}^{(s)\prime}\mathbf{z}_{ij}^{(s)} + \boldsymbol{\alpha}^{(s)\prime}\mathbf{c}_{j}^{(s)} + u_{ij}^{(s)} + v_{j}^{(s)}, \qquad (2.4.1)$$

where $\mathbf{x}_{tij}^{(s)}$ is a vector of time-variant covariates, with coefficient vector $\boldsymbol{\beta}^{(s)}$, such as indicators of the household's call history prior to t, the time and day of the current call,

information about the doorstep interviewer-householder interaction. The call history indicators include the number of calls made to the household until first contact was achieved and the number of intermediate noncontacts after first contact (i.e. some function of t), which are derived from *all* calls regardless of whether contact was made. The vector of time-varying covariates also includes an indicator of whether an appointment was made with the household at the previous call, which allows estimation of transition rates, that is, the probability that an appointment made at call $t-1$ is converted to cooperation at t. $\mathbf{z}_{ij}^{(s)}$ is a vector of time-invariant household covariates, with coefficient vector $\boldsymbol{\delta}^{(s)}$, such as type of accommodation, household in London. Time-invariant household characteristics include interviewer observations and census variables. The vector $\mathbf{c}_{j}^{(s)}$ includes time-invariant interviewer characteristics, such as gender, pay grade, with coefficient vector $\boldsymbol{\alpha}^{(s)}$.

Unobserved household and interviewer characteristics are represented respectively by random effects $u_{ij}^{(s)}$ and $v_{j}^{(s)}$. To allow for the possibility that some types of response outcome may have shared or correlated unmeasured influences, random effects at the same level are assumed to follow trivariate normal distributions: $\mathbf{u}_{ij} = (u_{ij}^{(1)}, u_{ij}^{(2)}, u_{ij}^{(3)}) \sim N(\mathbf{0}, \boldsymbol{\Omega}_u)$ and $\mathbf{v}_{j} = (v_{j}^{(1)}, v_{j}^{(2)}, v_{j}^{(3)}) \sim N(\mathbf{0}, \boldsymbol{\Omega}_v)$, where $\boldsymbol{\Omega}_u$ and $\boldsymbol{\Omega}_v$ are 3 × 3 covariance matrices. For example, similarity between the 'appointment' and 'other type of postponement' outcomes would be expected to lead to a positive correlation between $u_{ij}^{(2)}$ and $u_{ij}^{(3)}$ and between $v_{j}^{(2)}$ and $v_{j}^{(3)}$.

However, it is found that due to relatively small number of households with repeated outcomes of the same type there is not enough information in the data to estimate the household- and interviewer-level variances and covariances from the multinomial model (2.4.1) with outcome-specific household and interviewer random effects. Therefore, a second approach is proposed where outcome-specific loadings are used to overcome the estimation issue but to still allow for the effect of the household and interviewer unobservables to vary across outcomes. A simplified multilevel multinomial logit model for the log-odds of outcome s relative to outcome 4 (cooperation) may be written

$$\log\left(\frac{\pi_{tij}^{(s)}}{\pi_{tij}^{(4)}}\right) = \boldsymbol{\beta}^{(s)\prime}\mathbf{x}_{tij}^{(s)} + \boldsymbol{\delta}^{(s)\prime}\mathbf{z}_{ij}^{(s)} + \boldsymbol{\alpha}^{(s)\prime}\mathbf{c}_{j}^{(s)} + \lambda^{(s)}u_{ij} + \gamma^{(s)}v_{j}\,, \qquad (2.4.2)$$

where unobserved household and interviewer characteristics are represented respectively by normally distributed common random effects u_{ij} and v_j : $u_{ij} \sim N(\mathbf{0},\sigma_u^2)$ and $v_j \sim N(\mathbf{0},\sigma_v^2)$. These random effects have now outcome-specific coefficients or "loadings" $\lambda^{(s)}$ and $\gamma^{(s)}$ respectively, with $\lambda^{(1)}$ and $\gamma^{(1)}$ fixed to 1 for identification. Thus, although there is a set of unmeasured household and interviewer characteristics that affect the odds of all non-participation outcomes, their effects may differ across the three different survey outcomes. Outcome-specific loadings also allow the between-household and between-interviewer variance in the log-odds of non-participation to differ across outcomes. For example, the between-household variance is σ_u^2 for refusal (due to the identification constraint $\lambda^{(1)}$=1) and $(\lambda^{(2)})^2\sigma_u^2$ for appointments ($s = 2$). All other components in the model remain as in model (2.4.1).

Slightly surprisingly, after fitting model (2.4.2), no significant differences across the three interviewer-level random effect loadings are found, suggesting that unmeasured interviewer characteristics have similar effects on the log-odds of each type of non-participation outcome (the likelihood ratio test statistic for a test of the null hypothesis H_0: $\gamma^{(1)} = \gamma^{(2)} = \gamma^{(3)} = 1$ is 2.8 on 2 d.f., p=0.246). That is, there is no evidence for differential random interviewer effects on the three non-participation outcomes due to unobserved interviewer characteristics.

A simplification of model (2.4.2) with loadings on the interviewer random effect $\gamma^{(s)}$ (s =1,2,3) constrained to be equal across all three outcomes, may be written

$$\log\left(\frac{\pi_{tij}^{(s)}}{\pi_{tij}^{(4)}}\right) = \boldsymbol{\beta}^{(s)'}\mathbf{x}_{tij}^{(s)} + \boldsymbol{\delta}^{(s)'}\mathbf{z}_{ij}^{(s)} + \boldsymbol{\alpha}^{(s)'}\mathbf{c}_{j}^{(s)} + \lambda^{(s)}u_{ij} + v_j \,. \qquad (2.4.3)$$

This chapter uses model (2.4.3) as the final model and presents results in the following section.

The analysis file contains a record for each call that resulted in contact being made with the household. Each household may therefore contribute multiple records, up to a maximum of T_i, with their sequence of calls terminating in refusal, cooperation or the interviewer giving up (right-censored histories). Estimation of the models presented above is carried out using maximum likelihood as implemented in the aML software package (Lillard and Panis 2003). Where a closed form solution to the maximum likelihood function does not exist the residuals at each level are 'integrated

out' numerically using Gauss-Hermite quadrature. The number of quadrature points used is 16. Approximate standard errors are computed based on an approximation to the Hessian matrix.

To aid interpretation of the fitted model, predicted probabilities of each type of response outcome might be calculated for each value of a given covariate, holding constant the values of all other covariates in the model at their sample means. Population averaged probabilities might be obtained as follows: (i) take a large number M of random draws from the household and interviewer random effect distributions (based on the estimated random effect variances); (ii) calculate a set of predicted probabilities based on each set of generated random effect values and the estimated coefficients; and (iii) calculate for each outcome s the mean of the predicted probabilities $\hat{\pi}_{tij}^{(s)}$ across the M random effect values, where from (2.4.3)

$$\pi_{tij}^{(s)} = \frac{\exp\ \boldsymbol{\beta}^{(s)'}\mathbf{x}_{tij}^{(s)} + \boldsymbol{\delta}^{(s)'}\mathbf{z}_{ij}^{(s)} + \boldsymbol{\alpha}^{(s)'}\mathbf{c}_{j}^{(s)} + \lambda^{(s)}u_{ij} + v_j}{1 + \sum_{r=1}^{3}\exp\ \boldsymbol{\beta}^{(r)'}\mathbf{x}_{tij}^{(r)} + \boldsymbol{\delta}^{(r)'}\mathbf{z}_{ij}^{(r)} + \boldsymbol{\alpha}^{(r)'}\mathbf{c}_{j}^{(r)} + \lambda^{(r)}u_{ij} + v_j}, \qquad s = 1,2,3,$$

$$\pi_{tij}^{(4)} = 1 - \pi_{tij}^{(1)} - \pi_{tij}^{(2)} - \pi_{tij}^{(3)}.$$

To evaluate the model fit, likelihood ratio tests are used (Goldstein, 2010). This allows the comparison of nested models, for example, to evaluate if the addition of call record variables leads to a significant improvement in comparison to a simpler model without such variables.

2.4.3 Results

To help understand the process leading to cooperation or refusal, some descriptive statistics and preliminary modelling are initially presented. Table 2.4.1 illustrates the probability of each outcome at the first contact with the household by time of day and day of the week. At first contact, it may be assumed that the interviewer has little, if any, information about the household that might influence his/her calling behaviour. In particular, the first contact call is not affected by a previous appointment made. Table 2.4.1 shows that most first contacts are made on weekday afternoons, followed by weekday evenings and weekday mornings, with a clear decline in the number of contacts from the beginning to the end of the week for all times of the day.

Overall, 26% of all first contact calls results in a successful interview, 8% in refusal, 43% in an appointment made, and the remaining 24% in another type of postponement. The probability of immediate cooperation at the first contact call is highest (above 30%) for morning and afternoon calls at the beginning of the week, Monday and Tuesday, with a clear decline as the week progresses. The lowest cooperation rates are in the evening in particular towards the end of the week. On the other hand, householders are more likely to make an appointment with the interviewer if contact is made in the evening (above 45%) and this is for all days of the week but especially at the weekend. The probability of refusal and other forms of postponement are fairly stable at around 8% and 25% respectively by time of contact. It should be noted that only a few first contact calls are made at the weekend, in particular on Sunday. These findings are in line with previous literature (Purdon et al., 1999).

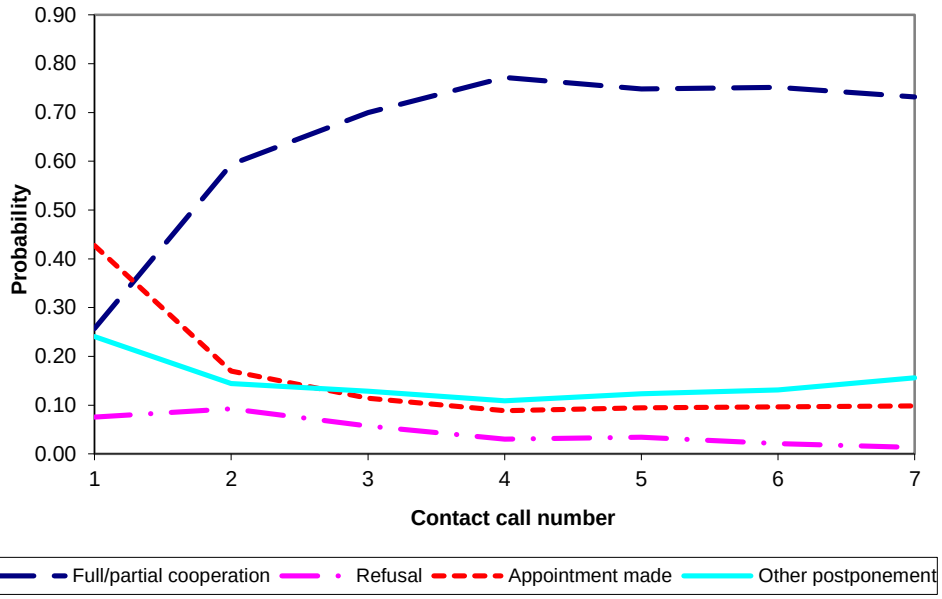
Table 2.4.1: Probability of each outcome at first contact, by day and time of call

		Cooperation	Refusal	Appointment made	Other postponement	Total number of first contact made	% of all first contacts
Monday	am	0.37	0.09	0.34	0.21	381	2.41
	pm	0.37	0.07	0.32	0.24	2162	13.70
	eve	0.25	0.08	0.48	0.20	1648	10.44
Tuesday	am	0.31	0.09	0.34	0.26	279	1.77
	pm	0.31	0.06	0.37	0.26	1919	12.16
	eve	0.23	0.08	0.49	0.21	1649	10.45
Wednesday	am	0.29	0.12	0.40	0.20	214	1.36
	pm	0.26	0.07	0.43	0.24	1544	9.78
	eve	0.20	0.08	0.48	0.24	1472	9.33
Thursday	am	0.28	0.09	0.39	0.25	212	1.34
	pm	0.22	0.08	0.42	0.28	1253	7.94
	eve	0.19	0.08	0.46	0.27	1138	7.21
Friday	am	0.23	0.12	0.39	0.27	151	<1.0
	pm	0.20	0.07	0.46	0.27	735	4.66
	eve	0.18	0.10	0.51	0.22	580	3.68
Saturday	am	0.26	0.05	0.43	0.27	109	<1.0
	pm	0.14	0.08	0.54	0.24	239	1.51
	eve	0.12	0.04	0.52	0.33	52	<1.0
Sunday	am	0.20	0.20	0.30	0.30	10†	<1.0
	pm	0.11	0.05	0.68	0.16	19†	<1.0
	eve	0.06	0.00	0.69	0.25	16†	<1.0
Total		0.26	0.08	0.43	0.24	15782	100

Morning (am): 0.00-12.00, Afternoon (pm): 12.00-17.00, Evening (eve): 17.00-0.00
† indicates cells with a sample size of less than 30

Since the first contact call is only indicative of the chances of achieving cooperation with a household this study now examines changes in the rates of the different outcomes across calls. Figure 2.4.1 shows the specific-outcome rates for the first seven contact calls. From contact call 7 onwards each outcome rate is based on few cases, if any, and so results for these contact calls are not presented here.

Figure 2.4.1: Specific-outcome rates by contact call number, allowing for repeated cooperation events

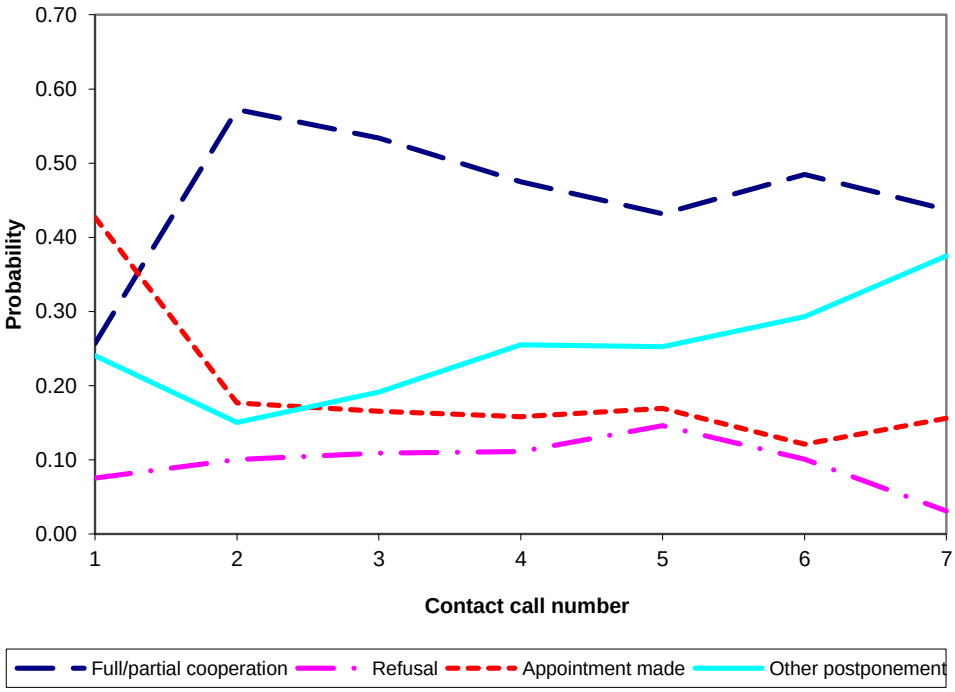


The chance of making an appointment is highest at the first contact call, when about 43% of calls end in an appointment made with the householder. It substantially decreases at second call to about 17% and then remains stable at around 10% for all subsequent calls. That is, after the second contact 1 in 10 visits are likely to finish in an appointment made with the householder. The cooperation rate is lowest at the first contact call (26%), increases sharply at the second to about 60% and then stabilises at just above 70% at the fourth and subsequent calls. The rise in the cooperation rate for calls 2 to 4 may be explained by the large number of appointments that were made at the early calls, in particular at the first call. It may be speculated that prior appointments are usually turned into successful interviews at the next call. The growth in the cooperation rate, even after initial contact calls, might also be explained by the presence of households with more than one person (multiple households). In these cases, the interviewer might seek to obtain cooperation from each household member at different

calls. The refusal rate is highest at the first and second contact call (at around 8%) to then quickly decrease towards zero as the number of contact increases. The behaviour of the refusal rate seems to indicate that people that are inclined to refuse do so in early calls. Other forms of postponement are relatively high at the first call (25%), then fall to just over 10% and continue to rise again steadily from call 4 onwards. Taken together, these patterns suggest that for later calls (from about call 4 onwards) the household either cooperates or the interviewer decides to postpone to another time or to stop calling, rather than continuing until receiving a refusal.

It may be argued, as mentioned before, that the results displayed in Figure 2.4.1 may be driven by the presence of multiple households in the sample, where cooperation from each household member could be obtained at different calls. Instead of looking at all contact calls, including cooperation occurring more than once to a household, it is possible to investigate the change in the rates of the different outcomes across calls *until first cooperation* is obtained, presented in Figure 2.4.2.

Figure 2.4.2: Specific-outcome rates by contact call number, until first time cooperation



Comparing the behaviour of the curves for appointments and other forms of postponement, Figure 2.4.1 and 2.4.2 show overall fairly similar results; for first time cooperation, however, the other postponement rate increases more rapidly with each additional contact, maybe indicating a hidden refusal. The cooperation rate across calls, and thus also the refusal rate, performs rather differently from call 2 onwards. Similarly to Figure 2.4.1 the cooperation rate increases sharply at the second contact (from 26% to 57%), but then decreases as the number of contacts increases. The pattern suggests that after the initial rise in the cooperation rate, possibly explained by the number of appointments made at the first contact, the likelihood of gaining first cooperation from the household reduces with each additional contact call. Likewise, the likelihood of refusal increases with the number of contacts made to the household.

Table 2.4.2: Estimated coefficients for the variable 'day and time of call' when included as a main effect in a multilevel multinomial logistic model controlling for household and interviewer characteristics

		Refusal $\hat{\beta}$ (ste)	**Appointment** $\hat{\beta}$ (ste)	**Other postponement** $\hat{\beta}$ (ste)
Monday	Morning	-0.149 (0.210)	-0.185 (0.125)	-0.019 (0.152)
	Afternoon	-0.449 (0.123)	-0.497 (0.075)	-0.054 (0.089)
	Evening	ref	ref	ref
Tuesday	Morning	-0.069 (0.201)	-0.341 (0.129)	0.040 (0.149)
	Afternoon	-0.707 (0.126)	-0.475 (0.076)	-0.094 (0.090)
	Evening	-0.217 (0.120)	-0.037 (0.072)	0.025 (0.086)
Wednesday	Morning	-0.215 (0.206)	-0.464 (0.133)	-0.584 (0.164)
	Afternoon	-0.677 (0.130)	-0.475 (0.078)	-0.168 (0.093)
	Evening	-0.353 (0.119)	-0.133 (0.073)	-0.011 (0.087)
Thursday	Morning	-0.821 (0.221)	-0.527 (0.133)	-0.389 (0.155)
	Afternoon	-0.397 (0.135)	-0.333 (0.082)	0.061 (0.096)
	Evening	-0.313 (0.125)	-0.142 (0.076)	0.090 (0.090)
Friday	Morning	-0.285 (0.254)	-0.359 (0.157)	0.017 (0.174)
	Afternoon	-0.438 (0.161)	-0.326 (0.097)	0.047 (0.112)
	Evening	-0.204 (0.155)	-0.150 (0.098)	-0.128 (0.116)
Saturday	Morning	-0.420 (0.286)	-0.407 (0.172)	-0.178 (0.195)
	Afternoon	-0.050 (0.239)	0.129 (0.147)	0.228 (0.171)
	Evening	-2.028 (0.787)	-0.337 (0.278)	0.186 (0.306)
Sunday	Morning	0.569 (0.691)	0.616 (0.412)	0.418 (0.517)
	Afternoon	-1.389 (0.679)	0.008 (0.396)	-0.844 (0.516)
	Evening	0.672 (0.684)	1.610 (0.350)	0.621 (0.450)

In order to investigate the influence of the day of the week and the time of the day on each possible outcome, it is convenient to recode the calling time variable reducing its 21 categories. To identify any reasonable pattern on this variable, its effect is examined in a multilevel multinomial model controlling for household and interviewer characteristics. The estimated coefficients for each category of the calling time variable are provided in Table 2.4.2. The sorted net effects on hazards (not shown here), together with the indicative findings of Table 2.4.1, suggest a quite different pattern for early week and late week, with Sunday more like the early part of the week and Saturday more like late week, especially Friday. In addition, the few calls made on Saturdays and Sundays made it necessary to merge these categories with other days of the week. These results informed the categorisation of the calling time variable in the final model (Table 2.4.3) which distinguishes six categories: early week (Sun-Tue) and late week (Wed-Sat) and morning, afternoon and evening.

Multilevel multinomial model

This section discusses the results from the final multilevel multinomial logistic model which includes time-varying call characteristics, fixed interviewer observations, household and interviewer characteristics, and household and interviewer random effects. The model aims to investigate the effect and usefulness of call record data and other paradata on the process leading to cooperation or refusal. Here this process is modelled across all contact calls. Of particular interest are the influences on a call outcome of the interaction between the interviewer and householder at the doorstep, of time-varying factors, such as number of previous calls, number of intermediate noncontacts after first contact was made, whether a prior appointment is made and of time-invariant interviewer observation about the household and neighbourhood. The model aims to control for all these factors. The model also controls for household information, primarily from the census, and interviewer characteristics that might be related to differential calling behaviours, in an attempt to adjust for the potential bias introduced by not fully randomised calling times for sample units (see section 2.3.3). It should be noted that the model discussed in this section includes an indicator if previously cooperation with the household has been achieved. For matter of completeness, the process leading to first time cooperation is also modelled for comparison. It is found that the results are very close to the final results presented in

this section (results are not shown). Parameter estimates of the final model are presented in Table 2.4.3.

Time-varying call characteristics

The inclusion of time-varying covariates into the model in comparison to a model with only census household level variables indicates a significant better fit (likelihood ratio test statistic is 2*14216, on 75 d.f., p=0.000), supporting results in Bates et al. (2008) who found that the inclusion of such variables 'greatly improve' models predicting nonresponse. The final model includes an indicator of whether previous contact was made with the household and the number of previous calls, distinguishing between contact and noncontact calls. It also controls for the number of times noncontact occurs after the first contact with the sample unit was achieved (intermediate noncontact). The previous contact indicator means that the coefficients of number of contact calls are interpreted as the effect on the different forms of non-participation of each additional call after the first call. This indicator is also included in the model to deal with some variables, such as previous appointment indicator, that are not defined at the first contact. It is observed from the model that the probabilities of refusal, appointment made and other forms of postponement decrease significantly with each additional call after first contact was made with the household, controlling for the other explanatory variables in the model. On the other hand, the odds of cooperation increase with each additional contact made. This is in line with previous research that report strong positive effects of having prior contact with the household on the propensity of an interview (Groves and Heeringa, 2006). This effect may indicate that keeping an ongoing interaction between the interviewer and the householder rather than seeking a quick decision on participation from the householder may be more likely to lead to a positive outcome. This would support the 'householder-interviewer interaction hypothesis' of Groves and Couper (1998, page 220). It could also indicate that interviewers are persistent in returning to a household if they feel they have a chance of a positive outcome. There is a (small) positive effect of the number of calls made until first contact on the probability of refusal, with negative effects on the other non-participation outcomes. This small effect on the likelihood of refusal may provide some evidence that households that are more difficult to reach may be more likely to refuse once contacted. The effect of the number of intermediate noncontact calls is positive for all three non-participation outcomes: the more intermediate noncontact calls are

made after first contact the more likely it is that the household refuses, makes an appointment or other form of postponement occurs. This may indicate that a noncontact could in fact be a hidden evasion or refusal, for example, due to fear of crime, which has been hypothesised in the literature (Groves and Couper, 1998; Stoop, 2005). However, the lack of a correlation between the noncontact and refusal processes identified in earlier research (Lynn et al. 2002; Nicoletti and Perachi, 2005; Steele and Durrant, 2011) has so far not provided much support for this hypothesis. A model controlling for the additional outcome of a noncontact at a call may provide further evidence for this phenomenon. Leaving a card or message behind is not found to affect the probabilities of any type of nonresponse.

Regarding the timing of the call, there is significant evidence that the outcome of the call may be affected by the time of day and the day of the week. For example, the likelihood of refusal is lower for afternoon and late week evening but higher for morning and early week evening calls. As one may expect, with a prior appointment the likelihood for refusal at the next contact call is greatly reduced. Transition rates are also calculated, i.e. the probability that an appointment made at the previous call is converted to cooperation at the current call. If the previous call results in an appointment the chances of experiencing cooperation at the next call is high (around 80%), and this is found to hold for any time of day. For comparison, without a prior appointment predicted probabilities for cooperation are below 60% for any calling time. The model controls for the case where an appointment is made and the following contact call results in another appointment or other form of postponement, such as a broken appointment or where the interviewer withdraws to try again later, which may indicate a potential lack of willingness to cooperate. Without a previous appointment made, the probability of an appointment at the current call is significantly higher (around 30%) than that for cases with a previous appointment (around 10%). Similarly, without a previous appointment, the probability of other forms of postponement at the current call is significantly higher (around 15%) than that for cases with previous appointment made (around 9%). These transition rates hold for any calling time. No significant interaction between previous appointment made and calling time is found.

It is important to observe that estimating causal effects of time-varying factors such as day and time of call would require randomisation of interviewers to different calling strategies. The model attempts to control for differences between these interviewer calling strategies and approximate the design that would be required for

estimating causal effects by including selected household and interviewer characteristics and a previous appointment indicator in the model. The effects of time of the call should be, however, interpreted with caution.

Of particular interest is the effect on the call outcome of what happens at the doorstep, especially the initial interaction between the householder and the interviewer. Survey organisations might be able to train interviewers to react accordingly to what happens at the doorstep or adequately schedule a re-call after some doorstep information has been gathered. In particular, the mode of contact appears relevant for the likelihood of gaining immediate cooperation: the chances of a refusal, making an appointment or other form of postponement are significantly lower if the contact is face-to-face rather than through an intercom system, a closed window or a closed door. This effect remains after controlling for potential interviewer observation effects about the area and household characteristics such as rural/urban households. Non face-to-face contact could indicate a potential fear of crime or a reluctance to talk to strangers which has been shown in other studies to lead to a higher refusal rate (Groves and Couper, 1998). If the householder asks at least one question, the chances of refusal, appointment or postponement are significantly reduced. Likewise, if the householder makes at least one positive or neutral comment as opposed to no comment, the odds of refusal or the interviewer withdrawing are much reduced while the odds of making an appointment increase. As would be expected, people who engage in a positive or neutral way with the interviewer (asking a question or making a comment), potentially expressing some interest in the survey, tend to cooperate more than those who do not. On the other hand, if the householder makes at least one negative comment, refusal, appointment and other postponements are much more likely than if no comment was made.

Characteristics of the person the interviewer talked to at the doorstep (based on interviewer observations) also seem to be useful in predicting the outcome of the call. For example, older people (60 and over) are less likely to refuse, make an appointment or postpone. A higher cooperation rate for older householders has been noted in other studies (Groves and Couper, 1996). If the person at the doorstep is female the call is more likely to result in an appointment or a postponement rather than cooperation, which may reflect a greater reluctance to speak to strangers or fear of crime among women. There is no gender difference in the immediate refusal behaviour.

Table 2.4.3: Estimated coefficients (and standard errors in parentheses) of multilevel multinomial logistic model including household and interviewer random effects

Variable (ref = Reference category)	**Categories**	$\hat{\beta}$ $(ste(\hat{\beta}))$ ***Refusal***	$\hat{\beta}$ $(ste(\hat{\beta}))$ ***appointment made***	$\hat{\beta}$ $(ste(\hat{\beta}))$ ***other postponement***
Constant		-3.074 (0.237)***	0.058 (0.119)	-0.389 (0.138)***
Call record variables (time variant)				
Previous contact indicator (ref =First contact)	Contact previously made	1.351 (0.117)***	-0.462 (0.066)***	-0.527 (0.077)***
Number of contact calls previously made	-	-0.352 (0.050)***	-0.534 (0.032)***	-0.427 (0.034)***
Number of non-contact calls made until first contact	-	0.098 (0.021)***	-0.099 (0.011)***	-0.191 (0.015)***
Number of intermediate non-contact calls after first contact was made	-	0.378 (0.032)***	0.285 (0.020)***	0.182 (0.024)***
Day and time of contact † (ref =Sun-Mon-Tue eve)	Sun-Mon-Tue am Sun-Mon-Tue pm Wed-Thurs-Fri-Sat am Wed-Thurs-Fri-Sat pm Wed-Thurs-Fri-Sat eve	0.048 (0.163) -0.485 (0.101)*** -0.221 (0.139) -0.391 (0.098)*** -0.295 (0.092)***	-0.257 (0.087)*** -0.495 (0.052)*** -0.466 (0.076)*** -0.383 (0.051)*** -0.148 (0.049)***	-0.023 (0.104) -0.134 (0.063)** -0.331 (0.090)*** -0.083 (0.061) -0.005 (0.058)
Previous Appointment Indicator † (ref =No prior appointment made)	Prior appointment made	-3.397 (0.113)***	-2.560 (0.060)***	-2.454 (0.075)***
Previous Cooperation Indicator † (ref =No prior cooperation achieved)	Prior cooperation achieved	-5.653 (0.261)***	-2.712 (0.078)***	-2.784 (0.093)***
How contact was made at doorstep (ref =Face-to-face)	Not face-to-face	1.633 (0.108)***	2.227 (0.061)***	1.808 (0.071)***
Question made by householder during the interviewer introductory conversation (ref =No question made)	At least one question made	-1.545 (0.079)***	-0.370 (0.040)***	-1.146 (0.051)***
Comment made by householder during the interviewer introductory conversation (ref =No comment made)	Positive/neutral comment made At least one negative comment made	-0.892 (0.142)*** 4.987 (0.129)***	0.389 (0.043)*** 1.207 (0.062)***	-0.970 (0.053)*** 2.083 (0.063)***

Age of main person the interviewer talked to (ref =60 and over)	Less than 16 16-34 35-59	0.896 (0.543)* 0.223 (0.121)* 0.335 (0.098)***	1.507 (0.258)*** 0.713 (0.061)*** 0.517 (0.053)***	4.715 (0.211)*** 1.229 (0.074)*** 0.603 (0.065)***
Gender of main person the interviewer talked to (ref =Male)	Female	-0.102 (0.068)	0.225 (0.035)***	0.119 (0.042)***
Interviewer Observations (time invariant)				
Type of accommodation (ref =Not house)	House	0.446 (0.103)***	0.607 (0.056)***	0.542 (0.067)***
House in a better or worse condition than others in area (ref =Better/ About the same)	Worse	0.271 (0.123)**	0.207 (0.067)***	0.201 (0.079)**
Household-level variables (time invariant)				
Preschool children present (ref =No)	Preschool children	-0.337 (0.116)***	0.154 (0.053)***	-0.070 (0.063)
Household type (ref =Single household)	Couple household Multiple household	0.374 (0.078)*** 0.291 (0.224)	0.262 (0.041)*** 0.093 (0.117)	0.290 (0.049)*** 0.197 (0.134)
Urban/rural indicator (ref =Urban)	Rural	-0.192 (0.115)*	-0.124 (0.060)**	-0.213 (0.074)***
Indicator if adults in employment (ref =No adults)	One or more adults	0.115 (0.091)	0.150 (0.048)***	0.359 (0.058)***
Educational attainment of Household Reference Person (ref =No educational attainment/ A levels, GCSEs)	First/Higher/College degree/Other attainment	-0.319 (0.086)***	-0.069 (0.042)*	-0.186 (0.050)***
Survey indicator (ref =EFS)	FRS GHS OMN NTS LFS	-0.213 (0.125)* -0.534 (0.113)*** -0.432 (0.118)*** -0.979 (0.111)*** -2.909 (0.151)***	-0.148 (0.075)* -0.117 (0.067)* -0.803 (0.071)*** -0.428 (0.064)*** -2.746 (0.099)***	-0.156 (0.087)* -0.111 (0.077) -0.449 (0.081)*** -0.421 (0.072)*** -3.327 (0.123)***
Interviewer-level variables (time invariant)				
Interviewer experience (ref = 9 years or more)	Less than 1 year 1 to 2 years 3 to 8 years	0.277 (0.139)** 0.187 (0.119) 0.178 (0.112)	0.019 (0.100) 0.011 (0.086) -0.029 (0.083)	0.140 (0.109) 0.121 (0.093) 0.088 (0.089)

Interviewer qualification (ref = Degree or postgraduate, other higher education)	A levels, GCSEs Qualifications below this level, no qualification	-0.229 (0.091)** 0.256 (0.217)	-0.002 (0.066) -0.350 (0.155)**	-0.055 (0.071) -0.322 (0.170)*
Can convince reluctant respondents (ref = Less confident)	More confident	-0.397 (0.119)***	-0.169 (0.081)**	-0.406 (0.089)***
Should persuade most reluctant respondent (ref = Strongly agree/agree)	Neither agree nor disagree Disagree/strongly disagree	-0.378 (0.153)** 0.382 (0.120)***	-0.055 (0.111) 0.130 (0.086)	-0.214 (0.120)* 0.114 (0.093)

The model is estimated using full information maximum likelihood. Where a closed form solution to the maximum likelihood function does not exist the residuals at each level are 'integrated out' numerically using Gauss-Hermite quadrature. The number of quadrature points used is 16. Approximate standard errors are computed based on an approximation to the Hessian matrix. The missing value categories have been suppressed to save space.

* significant at the 10% level
** significant at the 5% level
*** significant at the 1% level

† variable included in an interaction

Coding of time of call: am = 0.00-12.00, pm=12.00-17.00, eve= 17.00-0.00

Time invariant interviewer observations and household characteristics

Here, the effects of time invariant interviewer observations and household characteristics on nonresponse are investigated. The likelihood ratio test, comparing the fit of a model without interviewer observation and household characteristics to a model with, indicates the significant better fit of the more comprehensive model (the likelihood ratio test statistic is 2*258.31, on 33 d.f., p=0.000). It is important to remember that this research, in contrast with most previous research, investigates the probability of cooperation or refusal at a particular call and not the final response outcome. For example, for a certain subgroup in the population, the immediate cooperation rate at a particular call might appear lower than expected from the literature but due to appointments and other forms of postponements the final cooperation rate may be higher in line with expectation.

Interviewer observations on the household and neighbourhood are found to be useful in predicting the outcome of a call. Direct observations about the household as well as interviewer evaluations of the area are explored. Results from the previous section on the process leading to contact, show that householders living in a house rather than a flat are more likely to be contacted. The analysis here shows that those living in houses compared to those living in flats have higher chances of immediate refusal, although they also have higher chances of making an appointment which might result in future cooperation. The interviewer is also asked to judge the condition of the house and area. Living in a house that the interviewer reports to be in a worse condition than others in the area is associated with higher rates of refusal, appointment made and other postponements, as might be expected since socially deprived households have been found to be less likely to cooperate in other studies (Goyder, 1987). Physical barriers to access the household, such as a locked common entrance, locked gate or entry phone, and the presence of security devices, such as security staff, CCTV cameras or burglar alarm, are not found to affect the probabilities of refusal and appointment made relative to cooperation. However, a positive significant effect of these physical impediments on the likelihood of other form of postponement is found (results not presented here).

Some of the variables considered in the present study are available from both the census and the interviewer observation questionnaire, for example information on the presence of children and the household type. Census variables, where available, are included in the final model due to higher quality than interviewer reports. Other studies

without access to census variables may be able to include similar information based on interviewer observations. For households with pre-school children the immediate refusal and other postponement rate are lower. Such households are, however, more likely to request an appointment for a different time. This may be expected since, for example, households with children might be contacted relatively easily (Table 2.3.5), but it may not be convenient to participate in a survey in the presence of children in which case an appointment for another time may be made. Refusals, appointments and other postponements are more likely outcomes than cooperation in urban areas and for couple households. Households with at least one member in employment are more likely to postpone either making an appointment or otherwise. Households where the household representative has a high educational attainment are less likely to refuse, to make an appointment (at a marginal level) or to postpone, leading to a higher cooperation rate (see also Goyder, 1987). After controlling for household characteristics, such as type of household, and type of area, the London indicator is no longer significant. However, in the absence of other information, this indicator may be regarded as a proxy for household characteristics and useful to predict cooperation.

The model also allows for differences in cooperation and refusal across the six surveys. It is found the highest refusal, appointment and postponement rates for the EFS, a survey with a relatively high response burden due to the requirement to keep a diary and a long questionnaire. The lowest rates are achieved for the LFS, a less burdensome survey with a comparatively short interview. Further details on the differences between the surveys and an analysis of survey-dependent effects on ultimate contact and refusal rates can be found in Durrant and Steele (2009).

Time invariant interviewer characteristics

The effects of a range of time invariant interviewer characteristics on each outcome are now investigated. These characteristics include interviewer attributes, such as experience and qualification, and interviewer attitudes towards participation, such as confidence and persuasion. The inclusion of interviewer characteristics into the model in comparison to a model with only census household level variables and call record information indicates a significant better fit (likelihood ratio test statistic is 2*50.61, on 24 d.f., p=0.000). Several studies have explored the role of the interviewer in survey nonresponse and found that length of interviewer experience positively influence response rates (Couper and Groves, 1992; Pickery and Loosvelt, 2002; Hox and de

Leeuw, 2002; Durrant et al., 2010). This study also finds it to predict higher cooperation rates for more experienced interviewers. Interestingly, no effects of experience on appointments are found. Experience interviewers might achieve higher response rates by adopting certain strategies, such as appear trustworthy and friendly, adapt to the situation at the doorstep and react to the respondent, more efficiently than less experience interviewers (Morton-Williams, 1993). However, it is important to note that self-selection of interviewers might make it difficult to determine causation of length of interviewer experience. It may be likely that more effective interviewers (as judged by response rates) stay in their jobs for longer than those performing worse. The effect of interviewer experience might be then interpreted with caution. Regarding interviewers' qualification, there is evidence in favour of interviewers with an academic attainment below college degree, such as A levels or GCSEs, performing better (lower refusal rates) than those with a college degree or higher. Interviewers with low or no qualifications seem to be less likely to experience appointments or other form of postponements.

Some interviewer demographic characteristics, such as gender and age, were also investigated. Age of the interviewer is not found to affect the probabilities of any type of nonresponse; while there is some evidence that female interviewers, which represent 41% of the interviewer workforce, are less likely to get a refusal from the householder than their male counterpart. This gender effect on response rate was also observed by Hox and de Leeuw (2002). This study do not seek to interpret these effects on cooperation as demographic interviewer characteristics are usually largely out of the control of the survey agencies. A much interesting effect, which is not investigated in this research, would be looking at the interaction between these interviewer characteristics and sample members characteristics to examine, for example, whether homogeneity between interviewers and householders may result in higher cooperation rates (Durrant et al., 2010; Groves and Couper, 1998).

Interviewer attitudes toward cooperation or refusal seem to be good predictors of response. A strongly significant effect of the confidence of the interviewer and the attitude towards persuasion of reluctant respondents, both measured independently of the survey in question, are found. Interviewers who report more confidence in their ability to persuade reluctant respondents show a lower probability of refusal. Interestingly these interviewers also experience significantly less appointments and other forms of postponement. Interviewers who agree they should persuade reluctant respondents also have a lower refusal rate than those that disagree with the assertion.

No differences on making appointments are observed. This finding is in line with the literature indicating that interviewers with a positive attitude towards persuasion strategies and who, prior to the survey, are confident about their ability to obtain cooperation tend to attain higher response rates (de Leeuw et al., 1997; Groves and Couper, 1998; Hox and de Leeuw, 2002; Durrant et al., 2010; Blom et al., 2010).

Random household and interviewer effects

Table 2.4.4 presents the estimated household and interviewer random effect parameters from the final multilevel multinomial logistic regression model (2.4.3). The results show significant residual variation in the log-odds of a nonresponse outcome between households and between interviewers, after adjusting for all other covariates in the model. This implies that household and interviewer characteristics indeed play an important role on the response outcome at a particular call, as would be expected in line with previous research on response outcome (O'Muircheartaigh and Campanelli, 1999, Pickery and Loosveldt, 2002, Durrant et al, 2010). Comparing results from the previous section on the process leading to contact, the variation between interviewers in their non-participating outcome rates ($\hat{\sigma}_v^2 = 0.27$) is higher than the variation in their contact rates ($\hat{\sigma}_u^2 = 0.08$). This might provide some evidence that interviewer effects are more important for the process leading to cooperation. This might be due to the fact that this process depends much more on interviewer skills and behaviours and the interaction between the interviewer and the householder at the doorstep than the process leading to contact, which is more determined by timings and household characteristics.

Unmeasured interviewer characteristics, represented by v_j, have the same effect on the log-odds of each of the three forms of non-participation. That is, no indication of differential random interviewer effects on the three non-participation outcomes due to unobserved interviewer characteristics is found (see section 2.4.2). In other words, there is no evidence to support a hypothesis that particular interviewer characteristics are associated with certain outcomes, for example, that certain types of interviewers prefer making appointments. At the household level, however, there is evidence of differential effects of unmeasured household characteristics u_{ij} across the three outcomes (based on *t*-tests that the loadings for postponement and appointment are equal to 1: $t = 3.1$, p=0.002 for H_0: $\lambda^{(2)} = 1$ and $t = 5.1$, p=0.000 for H_0: $\lambda^{(3)} = 1$). While there is significant between-household variation in the log-odds of all forms of

non-participation, household effects are strongest for refusal and weakest for other postponement. As the loading $\lambda^{(1)}$ is fixed at 1 (for refusal), negative loadings for appointment, $\lambda^{(2)}$, and other postponement, $\lambda^{(3)}$, suggest that the household unobservables that are positively associated with refusal are negatively associated with both appointment and other postponement. In other words, households whose unobserved characteristics place them at high risk of refusal tend to be less likely to postpone by making an appointment or otherwise. This can be thought as a negative correlation between a household's refusal and postponement propensities, after adjusting for the covariates in the model.

Table 2.4.4: Estimated household and interviewer random effect parameters from the multilevel multinomial logistic regression model (standard errors in parentheses)

Parameter	**Estimate (Standard Error)**
Household common standard deviation σ_u	0.823 (0.132)***
Household random effect loadings $\lambda^{(s)}$	
$\lambda^{(1)}$ Refusal	1[a]
$\lambda^{(2)}$ Appointment made	-0.440 (0.149)***
$\lambda^{(3)}$ Other postponement	-0.880 (0.217)***
Interviewer common standard deviation σ_v	0.515 (0.029)***

a Constrained to equal 1
*** Significantly different from zero at the 1% level

2.5 Summary and implications for surveys practice

This chapter deals with nonresponse in sample surveys during the data collection process. The research presented here benefits from the availability of relatively rich paradata from six UK interview administrated household surveys. The first part of this chapter develops propensity models that predict the likelihood of contact in the field conditioning on household and interviewer characteristics. It explores the best times to contact different types of households, controlling for interviewer and area effects. The second part focuses on understating the process leading to cooperation or refusal. Using multilevel multinomial logistic analysis, it jointly models four different outcomes at each call using interviewer call record and observation data and controlling for household and interviewer influences. This chapter

presents the analysis of call record data in a multilevel modelling framework A single-level model might provide a first working model, but may underestimate standard errors of regression coefficients, in particular of higher-level variables. In addition to such technical advantages, multilevel models also provide substantive benefits. In particular, multilevel models offer conclusions that go beyond the interpretation of single-level models. For example they allow exploring the influence of unknown household and interviewer characteristics on contact and cooperation via estimated random effects. A summary of main results and potential implications for survey practice is presented as follows:

1. The results support earlier findings that weekday evenings and weekend daytimes are, on average, the best times to call to achieve contact. However, without a prior appointment, households contacted at those times, in particular early week (Sun-Tue) evenings, are more likely to refuse, book an appointment or postpone in other form than those contacted at other times. It is also found that best times to contact depend on household characteristics, especially those related to at home patterns. Differences in optimal contact times have been found e.g. by type of accommodation and the presence of children, pensioners or unemployed persons. A call made at a time previously agreed through a booked appointment is most likely to lead to a successful interview for every time of the day and day of the week.
2. Interviewer observations about a household and neighbourhood, for example on the type and condition of the house and the presence of children, are useful for predicting the likelihood of contact and cooperation. Some interviewer observation variables are predictive of contact and cooperation before and after controlling for additional information about a household (from the census in the present study). These observations might be used as proxies for census information that is usually unavailable.
3. There is significant evidence that time-varying call record information, such as features of the call history and of the current call, play a key role in predicting contact and the outcome of each future call after contact was made. Of particular interest for survey agencies are interviewer strategies on establishing contact and gaining cooperation. The contact model shows some significant effects of such strategies, for example the probability of contact is higher at the next call if the interviewer left a card or message at a previous call. Regarding

cooperation, characteristics of the doorstep interaction process between the interviewer and the householder, such as how contact was established and whether the householder asked questions or made comments, are very relevant.

4. The multinomial model shows that controlling for all other variables in the model, the more contact calls is made the higher the odds of cooperation. This may provide some evidence that keeping in contact with the household may increase the chances of a successful interview. Rather than pressing for an immediate cooperation the interviewer may be advised to keep the conversation and the contact with the household going, for example by making an appointment for another time (Groves and Couper, 1998).
5. Area-level variables, geographical location and type of area are found predictive of contact before controlling for other household and calling variables, but they are not significant in the final model. Therefore, in the absence of additional information, area characteristics might be regarded as proxies for household characteristics and useful for predicting contact. Similarly, a London indicator is found predictive of cooperation before controlling for household characteristics, but it is not significant in the final model.
6. Significant effects of interviewer characteristics on contact and cooperation are observed. Important in explaining interviewer differences in contact rates are pay grade, qualifications and age. Interviewer experience is not found to be important on predicting contact after controlling for these factors. However, it is useful on predicting the likelihood of cooperation; more experience interviewers are likely to obtain higher cooperation rates. There is evidence that some interviewers may be more effective in establishing contact at certain times, which may indicate better judgement of when best to call. There is little empirical support for the hypothesis that some interviewers are more successful in establishing contact with more difficult households, such as single households. Strong effects of interviewer confidence and attitude towards persuasion of reluctant respondents are found in the multinomial model for cooperation; if interviewers express confidence in their abilities to convince reluctant respondents and agree they should persuade reluctant respondents, they are likely to achieve higher cooperation rates.
7. Some evidence for differential effects of fixed interviewer characteristics on refusal, appointment made and other forms of postponement is found. For

example, interviewer experience, although an important predictor of refusal, does not seem to impact on the likelihood of appointments or postponements. Interviewer confidence, on the contrary, impacts on all three non-participation outcomes.

8. Unmeasured interviewer characteristics have a significant effect on contact and cooperation. However, the variation between interviewers in their cooperation rates is higher than the variation in their contact rates. In the model for cooperation, no evidence for differential effects due to unmeasured interviewer characteristics on the three non-participation outcomes is found, i.e. the influence of the interviewer random effect is the same across the three non-participation outcomes.
9. At the household level, the multinomial model shows evidence of differential effects of unmeasured household characteristics across the three non-participating outcomes: refusal, appointment made and other postponement. Negative loadings for postponement outcomes suggest that household unobservables that are positively associated with refusal are negatively associated with both appointment and other postponement.

As discussed in the previous sections, the available data are based on a non-random allocation of calling times to households. The models attempt to control for household and interviewer characteristics likely to be associated with the interviewer decision on when to call. Nevertheless, as it is possible that the calling time may depend on unmeasured household and interviewer characteristics, the effects of day and time of the call should be interpreted with caution and inferences about possible causal effects of finding should be avoided.

The results have a number of potential implications for survey practice. They may inform the design of efficient and effective calling behaviours and follow-ups as well as responsive survey designs to increase response rates and to potentially reduce nonresponse bias. The type of models presented may be used to predict the likelihood of contact or cooperation at the next call, conditioning on information known to the survey organisation or interviewer at each point in time - even in the absence of information like here from the census. For example, an interviewer or survey agency may be able to observe hints for a potential refusal early on, before a hard refusal occurs. It might be then possible to intervene to avoid a refusal, for example, by offering a

higher incentive or by sending a more experienced interviewer. These models may also be used to estimate response propensities from sample units to be employed for adjustment and estimation at the data analysis stage (see Chapter 3). The focus of this research is on face-to-face surveys but some findings may also apply to telephone surveys.

The research highlights the benefits of prior information about sample units for improving prediction of contact and cooperation, and survey agencies should exploit possibilities of data linkage to boost information available about each household or area. Such additional information may come from the sampling frame, registers or administrative data, as well as previous waves in the case of a longitudinal study - available prior to data collection. The availability of such additional data may depend on the country and some restrictions on data linkage may apply due to confidentiality and data disclosure concerns. The analyses also highlight the relevance of call record information and interviewer observations (paradata) captured during data collection to inform the process leading to contact and cooperation. These variables could be used as proxies of household characteristics if, for example, census data are not available. This has also implications for survey organisations that need to carefully consider which type of paradata should be recorded at each call, such as outcome of the call, doorstep interactions with the householder and interviewer observations about the household and neighbourhood. They also need to assess how best to collect such data, including interviewer training.

The significant interviewer effects in predicting contact imply that survey agencies may have a greater choice than previously thought regarding how best to contact a household, rather than, as was hypothesised in Purdon et al. (1999), simply decisions on the timing of calls. For example, certain interviewers may be allocated to more difficult times or cases – at least within fieldwork constraints such as travelling times and costs. It may also be advantageous for the survey organisation to be aware of other time commitments of interviewers; for example interviewers who have only a limited capacity to make evening and weekend calls may need additional support or may be allocated certain cases or areas.

Chapter 3

Weighting adjustment for clustered nonresponse

3.1 Introduction

To produce more accurate estimates of population characteristics in the presence of nonresponse weighting adjustment is often carried out to reduce nonresponse bias in estimates from sample surveys (e.g. Little, 1986, 1988; Särndal and Lundström, 2005). As discussed in Chapter 1, a commonly used weighting technique is inverse probability weighting. This technique consists of deriving response propensities from sample units under a model and then using the inverse of these estimated probabilities as the adjustment weights (e.g. Ekholm and Laaksonen, 1991; Iannaccliione, 2003). The key to effectively model these response probabilities is the availability of auxiliary information for both the respondents and the nonrespondents to the survey. Information about the population distribution and other paradata, such as information on the interviewers, might also be used in the model. Adjustment weights may be combined with sampling weights for a joined treatment of nonresponse and sampling.

Most discussions of inverse probability weighting assume that responses for different units are independent (e.g. Cao et al., 2009; Kim and Kim, 2007). It is not uncommon in surveys, however, for nonresponse to be correlated within clusters. Chapter 2, for example, shows that interviewers are differentially successful at interviewing sample households leading to interviewer (cluster) differential response rates. The effect of observed and unobserved interviewer's characteristics on response may result in correlation between the likelihood to participate of different households approached by the same interviewer. For example, Durrant et al (2010) found that the likelihood of refusal is higher if the interviewer has a college degree but the householder does not, and it is highest for the case where the interviewer has only a low or no educational attainment and the householder has a professional degree of some form. Thus, correlated nonresponse might be observed due to the clustering of households

within interviewers. Another example might be a two-stage cluster sample with geographical areas as primary sampling units and households as secondary sampling units. If nonresponse of households depends on unmeasured area-level characteristics, nonresponse intra-cluster correlation may occur simply because of the heterogeneity between clusters used for sampling.

This chapter investigates how to construct inverse probability weights, when response is clustered and cluster membership is observed for both responding and nonresponding units, as is the case when the clusters are defined by interviewer workloads or they define a stage in a multi-stage sampling design. One established approach is to use such clusters (or homogeneous sets of clusters) as weighting adjustment cells (e.g. Little, 1986), where the implicit model is that response probabilities vary just by cell and may be estimated by the cell-level response rates. Weighting cell adjustment increases the weights of the respondents by the same amount in each cell so that the sum of the adjustment weights of the respondents equals the sum of the sampling weights of the complete sample within each cell. This chapter considers the more general setting when auxiliary information at the sample level is available and include other variables in addition to cluster membership. A natural model for nonresponse, given such auxiliary information, is a multilevel model (as discussed in Chapter 2), where clustered nonresponse is captured via random effect terms. Peytchev (2011), for example, used information on interviewers as auxiliary data in the estimation of response propensities to adjust for unit nonresponse. However, Peytchev (2011) only used a single-level logistic regression approach ignoring the clustering in the data. This chapter investigates how to construct inverse probability weights based on multilevel models and assess to what extent these inverse probability weights result in more efficient estimates than those obtained by using simpler models that ignore the clustered data.

Yuan and Little (2007) proposed several methods to correct for unit nonresponse bias in a two-stage clustered survey. These methods were based on a random effects model for the survey variable and thus fall outside the class of weighting methods considered here, which aim to model response propensities. This chapter, however, makes use of the concept of cluster-specific non-ignorable (CSNI) nonresponse proposed by Yuan and Little (2007) to describe the case when nonresponse may depend on unobserved cluster random effects which may be correlated with the survey variables. Following the example above on educational

attainment of interviewers and householders, if household educational level is correlated with the survey variables of interest then the dependence of nonresponse on interviewer characteristics may result in a CSNI nonresponse mechanism. The cluster-specific non-ignorable nonresponse has also been discussed, at least implicitly, by Little and Rubin (2002, Example 6.24), Shao (2007) and Yuan and Little (2008). The CSNI condition is weaker than the usual missing at random (MAR) condition, where the probability of response is independent of the survey variables but may depend on other observed auxiliary variables. The MAR assumption is conventionally assumed if inverse probability weighting is to correct for bias (e.g. Tsiatis, 2006, p.146). A key aim of this chapter is to construct weights which exploit the auxiliary information on cluster membership and other variables to correct for bias under CSNI, not just MAR.

This chapter considers three 'standard' ways of constructing inverse probability weights, including the use of multilevel models as in Chapter 2 and a marginal model that ignore the clustering structure of the data, and a new proposed approach using conditional logistic regression. It also presents variance estimators for each adjustment weighted estimator, assuming weights are treated as fixed. Skinner and D'Arrigo (2011) proposed a more complex variance estimator that accounts for the fact that the weights are estimated. The properties of the alternative weighted estimators and associated variance estimators are investigated through a simulation study. Results from an empirical application using data from the Expenditure and Food Survey 2001 are also presented.

The chapter is organised as follows. The basic estimation and modelling framework is set out in section 3.2. The different ways of constructing inverse probability weights for cluster nonresponse is presented in section 3.3. Variance estimation is considered in section 3.4. A simulation study is presented in section 3.5. Section 3.6 shows an empirical illustration and some final discussion follows in section 3.7.

3.2 Estimation and modelling framework

Consider a finite population $U = \{(i,j) \mid i = 1,...,N, j = 1,...,M_i\}$, with the jth unit in the ith cluster labelled (i,j), from which a probability sample $s = \{(i,j) \mid i = 1,...,n, j = 1,...,m_i\} \subset U$ is drawn with a given sampling design. Suppose

that $\pi_{ij} = \Pr\ (i,j) \in s$, the probability of selection of (i,j) under the sampling design, is known and non-zero for each $(i,j) \in s$. Denote the population size $M = \sum_1^N M_i$ and the sample size $m = \sum_1^n m_i$. It is of interest to estimate the population total of a generic survey variable y , namely

$$T_y = \sum_{(i,j)\in U} y_{ij}$$

where y_{ij} denotes the value of y for the (i,j) unit of U . Note that many other parameters may be expressed as a function of such totals and estimated by this function of the corresponding estimated totals.

Let R_{ij} denote the binary response indicator variable, which is defined for all units $(i,j) \in U$, irrespective of which sample s is selected, such that

$$R_{ij} = \begin{cases} 1 & \text{if unit } (i,j) \text{ responds} \\ 0 & \text{if unit } (i,j) \text{ does not respond} \end{cases}$$

It is assumed that R_{ij} is a characteristic of the units in the population, and therefore that its values cannot change as a function of which sample s is selected ('stable' nonresponse in the terminology of Rubin, 1987, page 30). Thus, it is supposed that sampling and nonresponse are unconfounded, i.e. the sample is selected independently of the population values of R_{ij}.

Suppose that R_{ij} , a $1 \times k$ vector of auxiliary variables x_{ij} and the cluster membership indicator i are observed for all units in s , but that y_{ij} is only observed for respondents, i.e. for units in $\{(i,j) \in s : R_{ij} = 1\}$.

The primary focus of this chapter is on the inverse-probability weighted estimator of T_y given by

$$\hat{T}_y = \sum_{(i,j,)\in s} d_{ij}\hat{q}_{ij}R_{ij}y_{ij}\,, \tag{3.2.1}$$

where $d_{ij} = \pi_{ij}^{-1}$ is the design weight and $\hat{q}_{ij}$ is a non-response weight, representing an inverse estimated response probability, to be discussed in section 3.3. The estimator in (3.2.1) is called the two-phase nonresponse adjusted estimator in Särndal and Lundström (2005, equation 6.3).

This chapter also considers the so called two-phase generalized regression estimator (Särndal and Lundström, 2005, equation 6.4)

$$\hat{T}_{yreg} = \hat{T}_y + (\hat{T}_{xs} - \hat{T}_x)\hat{\lambda}, \tag{3.2.2}$$

where

$$\hat{T}_{xs} = \sum_{(i,j,)\in s} d_{ij}x_{ij}\,,\ \hat{T}_x = \sum_{(i,j,)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij},\ \text{and}\ \hat{\lambda} = (\sum_{(i,j,)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij}^T x_{ij})^{-1} \sum_{(i,j,)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij}^T y_{ij}\,,$$

introduced by Cassel et al. (1983).

In order to construct the nonresponse weights $\hat{q}_{ij}$ and to assess the properties of the estimators of T_y, a modelling framework ξ for the generation of the R_{ij} and y_{ij} is introduced. It is supposed earlier that sampling and nonresponse are unconfounded, that is that the distribution of the R_{ij} does not depend on the sample outcome s. More generally, it is assumed sampling is non-informative in the sense that the distribution of (R_{ij}, y_{ij}) implied by ξ does not depend on s.

The basic parametric model considered for R_{ij}, unconditional on y_{ij}, is

$$\Pr(R_{ij} = 1 \mid u_i) = h(x_{ij}\beta + u_i)\ ,\ u_i \sim N(0,\tau^2), \tag{3.2.3}$$

where u_i denotes a random cluster effect which captures the response intra-cluster correlation, $h(.)$ is a specified inverse link function, such as the inverse logit function, and the $k \times 1$ vector β and τ^2 are parameters. The R_{ij} are assumed mutually independent conditional on the u_i. This research only considers estimation in the case when the number of respondents in each cluster is non-zero. Yuan and Little (2007) commented on ways in which biased estimation can arise when this is not the case. For example, they noted that, if some sampled clusters do not have any respondents, their reweighted random-effects model based approach produces biased estimates as it ignores these clusters.

In addition to the random effects model (3.2.3), the implied marginal model is also considered:

$$\Pr(R_{ij} = 1) = g(x_{ij}\beta), \tag{3.2.4}$$

where $g(x_{ij}\beta) = E_{\xi}\left[h(x_{ij}\beta + u_i)\right]$ and the expectation is taken across the distribution of the random effect u_i. Note that the random effect will induce a correlation between R_{ij} and R_{ik} for $j \neq k$ in this model.

Response Mechanisms

For the response mechanism, two principal assumptions regarding the relation between R_{ij} and y_{ij} are considered. Nonresponse is said to be missing at random (MAR) if the R_{ij} and y_{ij} are mutually independent, that is $\Pr(R_{ij} = 1 \mid y_{ij}) = \Pr(R_{ij} = 1)$, given that x_{ij} is treated as fixed characteristics of the units in the population irrespective of which sample s is selected. The mechanism is said to be cluster-specific nonignorable nonresponse (CSNI), following Yuan and Little (2007), if model (3.2.3) holds and the R_{ij} and y_{ij} are independent conditional on the u_i, that is $\Pr(R_{ij} = 1 \mid y_{ij}, u_i) = \Pr(R_{ij} = 1 \mid u_i)$, again holding the x_{ij} fixed.

To illustrate and motivate the CSNI assumption, suppose y_{ij} obeys a linear multilevel model

$$y_{ij} = x_{ij}\lambda + v_i + \varepsilon_{ij}, \tag{3.2.5}$$

where v_i and ε_{ij} are nested random effect terms with zero means, such that the R_{ij} are conditionally independent of the v_i and ε_{ij} given the u_i and, furthermore, u_i is conditionally independent of the ε_{ij} given the v_i. Then, when both models (3.2.3) and (3.2.5) hold, nonresponse is MAR when u_i and v_i are independent and CSNI otherwise. The principal relevance of this chapter is to cases when CSNI holds but MAR does not. The key motivating application arises when both nonresponse and the survey variable exhibit clustering, which may be represented by the kind of joint cluster effect model for (R_{ij}, y_{ij}) in (3.2.3) and (3.2.5), where the cluster effects display correlation (after controlling for observable x_{ij}). For example, when clustering is by geography, correlation between area-level response rates and area means of the survey variable may be induced by a common correlation with average area-level income (which is not available as an x_{ij} variable).

3.3 Construction of nonresponse weight

This section considers the construction of the nonresponse weight $\hat{q}_{ij}$ used in the estimators in (3.2.1) and (3.2.2), when model (3.2.3) holds. It first considers three 'standard' options and then proposed a new approach using conditional logistic regression.

(i) *response propensity weights* (Little, 1988): the inverse link function $g(.)$ in the marginal probability $\Pr(R_{ij}=1)$ in (3.2.4) is assumed known and the weights are set to be $\hat{q}_{ij}^{M} = g(x_{ij}\hat{\beta}^{M})^{-1}$, where $\hat{\beta}^{M}$ is obtained, for example, by maximum likelihood estimation (MLE) under the working model of independent observations.

(ii) *weights based on predicted random effects*: set $\hat{q}_{ij}^{RE} = h(x_{ij}\hat{\beta}^{RE} + \hat{u}_{i}^{RE})^{-1}$, based on the random effects model in (3.2.3), where $\hat{\beta}^{RE}$ and the $\hat{u}_{i}^{RE}$ (and implicitly $\hat{\tau}^{2RE}$) might be predicted using an approximate ML method, such as in Diggle et al. (2002, p.174).

(iii) *weights based on estimated fixed effects*: set $\hat{q}_{ij}^{FE} = h(x_{ij}\hat{\beta}^{FE} + \hat{u}_{i}^{FE})^{-1}$ as in (ii), but where the u_i in (3.2.3) are now treated as unknown parameters (fixed effects, i.e. treating cluster as another explanatory variable with number of categories equal to the number of clusters appearing in the sample) and $\hat{\beta}^{FE}$ and the $\hat{u}_{i}^{FE}$ are MLEs. One advantage of this approach compared to (ii) when $h(.)$ is the inverse logit function is that it avoids numerical integration in the computation.

Skinner and D'Arrigo (2011) presented theoretical reasons why each of the above options may not correct adequately for bias from CSNI nonresponse when the m_i may be small. They proposed an alternative conditional logistic regression approach for this case, designed to remove the dependence of the weighting method on the random effects. The basic idea is to construct the weight as $\Pr(R_{ij}=1 \mid R_{i+})^{-1}$, where $R_{i+} = \sum_{j=1}^{m_i} R_{ij}$ is the number of respondents in cluster i. It may be shown (e.g. Agresti, 2002, p.251) that when model (3.2.3) holds and $h(.)$ is the inverse logit function,

$$\Pr(R_{ij}=1 \mid R_{i+}) = \frac{\sum_{\mathbf{r}_i \in B_{1ij}} \exp(\sum_{j=1}^{m_i} r_{ij} x_{ij} \beta)}{\sum_{\mathbf{r}_i \in B_{2i}} \exp(\sum_{j=1}^{m_i} r_{ij} x_{ij} \beta)} \quad , \tag{3.3.1}$$

where $\mathbf{r}_i = (r_{i1},...,r_{im_i})$ denotes the vector of observed response indicator values in cluster i, B_{1ij} represents the set of possible values of $\mathbf{r}_i$ where r_{ij}=1 and $\mathrm{r}_{i+} = R_{i+}$, i.e. $B_{1ij} = \{\mathbf{r}_i : r_{ij}=1, \mathrm{r}_{i+} = R_{i+}\}$ and B_{2i} denotes the set of possible values of $\mathbf{r}_i$ where $\mathrm{r}_{i+} = R_{i+}$, i.e. $B_{2i} = \{\mathbf{r}_i : \mathrm{r}_{i+} = R_{i+}\}$. The absence of the u_i in (3.3.1) arises from the sufficiency of R_{i+} for u_i. In practice, β is unknown and it is proposed to set $\hat{q}_{ij}^{CML} = \Pr(R_{ij}=1 \mid R_{i+} = r_{i+}; \beta = \hat{\beta}^{CML})^{-1}$, where $\hat{\beta}^{CML}$ is obtained by conditional ML (e.g. Agresti, 2002, p.496; see also Skinner and D'Arrigo, 2011).

The conditional logistic approach is closer to the fixed effects than the random effects approach in the sense that, given β, the weights in cluster i depend only on the R_{ij} in cluster i and they are not shrunk to a cluster average using outcomes from other clusters. In the special case when $x_{ij} = x_i$ and $x_{ij}\beta + u_i$ is replaced by u_i, since x_i is effectively confounded with u_i, both the conditional logistic and fixed effects weights reduce to m_i / R_{i+} (note that the sizes of B_{1ij} and B_{2i} are $\binom{m_i - 1}{R_{i+} - 1}$ and $\binom{m_i}{R_{i+}}$ respectively), the inverse response rate in cluster i, a traditional choice of weight with clustered survey data (Yuan and Little, 2007). Compared to the random effects approach, the conditional logistic approach has the advantage that it does not depend on assumptions about the distribution u_i nor about the relation of u_i to x_{ij}. On the other hand, it does depend on the assumption that h(.) is the inverse logit function in order that (3.3.1) holds and is free of u_i. Note that, since it was assumed that the sampling and nonresponse are unconfounded, design weights have not been incorporated in either the conditional probability in (3.3.1) or the construction of $\hat{\beta}^{CML}$.

The properties of the alternative weighted estimators of T_y denoted by $\hat{T}_y^M$, $\hat{T}_y^{FE}$, $\hat{T}_y^{RE}$ or $\hat{T}_y^{CML}$ when $\hat{q}_{ij} = \hat{q}_{ij}^M$, $\hat{q}_{ij}^{RE}$ or $\hat{q}_{ij}^{CML}$ in (3.2.1), and similarly

$\hat{T}^{M}_{yreg}, \hat{T}^{FE}_{yreg}, \hat{T}^{RE}_{yreg}$ or $\hat{T}^{CML}_{yreg}$ when the generalized regression estimator (3.2.2) is considered, are investigated through a simulation study in section 3.5.

3.4 Variance estimation

In the case of stratified selection of clusters, an approximated variance estimator of $\hat{T}_y$, treating adjust weights $\hat{q}_{ij}$ as fixed, may be written as (e.g. Stukel et al., 1996):

$$v = \sum_{h=1}^{H} \frac{n_h}{(n_h - 1)} \sum_{i \in s_h} (c_{i+} - \bar{c}_h)^2 \tag{3.4.1}$$

where $c_{i+} = \sum_{j=1}^{m_i} c_{ij} = \sum_{j=1}^{m_i} d_{ij} R_{ij} \hat{q}_{ij} y_{ij}$, $\bar{c}_h = n_h^{-1} \sum_{i \in s_h} c_{i+}$ and s_h denotes the set of n_h clusters drawn in stratum h, for $h = 1,...,H$ (it is assumed that $n_h \geq 2$ for each h). This effectively assumes that the c_{i+} may be treated as independent and identically distributed within strata, which may be a reasonable approximation for many sampling schemes where clusters are selected as primary sampling units (PSUs) and the fraction of PSUs selected in each stratum is small and when nonresponse is independent between clusters.

Skinner and D'Arrigo (2011) outlined a variance linearization approach that allows for variability on the estimated weights for the CML case. This approach will not be discussed in this thesis.

3.5 Simulation study

3.5.1 Description of the study

A simulation study is now carried out to illustrate the properties of the four weighted point estimators in section 3.3 and the variance estimator presented in the previous section.

Six finite populations with $N = 200$ and $M_i = M = 10$ are constructed, where the values of x_{ij}, R_{ij} and y_{ij} are generated, respectively, from:

$x_{ij} = (1, x_{1ij})$, $x_{1ij} \sim N(2,1)$, truncated below by 0 and above by 4,

$R_{ij} \sim$ model (3.2.3) with $h(.)$ the inverse logit function, where $\beta = (\beta_0, \beta_1)^T, \tau^2 = 1$,

$y_{ij} \sim$ model (3.2.5) with $\lambda = 5$, $\varepsilon_{ij} \sim N(0,1)$ and $v_i = \alpha_i + \delta u_i$, where $\alpha_i \sim N(0,1)$.

Since α_i, u_i and ε_{ij} are generated independently, nonresponse is MAR if $\delta = 0$ and CSNI otherwise. The six finite population are created following six possible sets of values for the parameters $\beta = (\beta_0, \beta_1)^T$ and δ, representing different missing data mechanisms, like this

(i) MCAR: $(\beta_0, \beta_1) = (1,0)$, $\delta = 0$

(ii) MAR: $(\beta_0, \beta_1) = (0,0.5)$, $\delta = 0$

(iii) CSNI1: $(\beta_0, \beta_1) = (1,0)$, $\delta = 5$

(iv) CSNI2: $(\beta_0, \beta_1) = (0,0.5)$, $\delta = 5$

(v) CSNI3: $(\beta_0, \beta_1) = (1,0)$, $\delta = 1$

(vi) CSNI4: $(\beta_0, \beta_1) = (0,0.5)$, $\delta = 1$

Note that mechanism (i) is described as missing completely at random (MCAR) since R_{ij} is independent of both y_{ij} and x_{ij}. The values of (β_0, β_1) above are chosen so that the overall response rate is approximately 70% and the nonresponse is generated independent $(\beta_1 = 0)$ or dependent $(\beta_1 = 0.5)$ of covariates. The alternatives values of δ determine the strength of the intra-cluster correlation for the values of y_{ij}. The intra-cluster correlation of y_{ij} defined by

$$\sigma_v^2 \Big/ \sigma_v^2 + \sigma_x^2 \lambda^2 + \sigma_\varepsilon^2 \;=\; 1 + \delta^2 \;\Big/\; 27 + \delta^2 \;,$$

is 0.037 in the MCAR and MAR cases $(\delta = 0)$, 0.5 in the CSNI1 and CSNI2 cases $(\delta = 5)$, and 0.07 in the CSNI3 and CSNI4 cases $(\delta = 1)$, designed to reflect a realistic range of possible values.

Two sampling designs are applied to these populations: (a) simple random cluster sampling with $n = 50$, $m_i = M = 10$; (b) two stage sampling, with simple random sampling at each stage with $n = 50$, $m_i = 5$. Each sampling scheme is repeated 1000 times for each population. New values of the response indicator R_{ij} are generated along

with the new samples, while other finite population values are kept fixed. Any samples for which $R_{i+} = 0$ for some i are rejected. The estimates of T_y and associated variance estimators for the following four weighting approaches are computed:

(i) M: marginal model (3.2.4)

(ii) RE: random effects model (3.2.3)

(iii) FE: fixed effect model (3.2.3) but with random effects treated as unknown parameters

(iv) CML estimated: weights based on (3.3.1) using conditional ML to estimate β^{RE}, where β^{RE} defines the true model when (3.2.3) holds.

To help understand the impact of estimating β^{RE} by $\hat{\beta}^{CML}$, $\hat{T}_y^{CML}$ with $\hat{\beta}^{CML}$ replaced by β^{RE}, referred as $\tilde{T}_y^{CML}$, is also included (i.e. 'CML true parameter' in Tables 3.5.1 and 3.5.2).

3.5.2 Results of the study

Tables 3.5.1 and 3.5.2 show summary statistics of weighted estimates of the total T_y for the different approaches in section 3.3 and for the alternative missing data mechanisms and choices of (n, m_i) above. The relative bias reported in the tables is the mean of the estimated total across the 1000 samples less the true population total, divided by this population total. The relative standard error (SE) is the standard deviation of the estimated total across the 1000 replications divided by the true population total. The relative root mean squared error (RMSE) is the square root of the average squared deviation of the estimated total from the true population total over the 1000 samples divided by the true population total.

Bias and standard error properties of the adjusted point estimate

No evidence of bias is observed in $\hat{T}_y^M$ under MAR or MCAR, as expected from Skinner and D'Arrigo (2011); however, this estimator is significantly biased under the CSNI mechanism. The bias of $\hat{T}_y^M$ decrease when a lower intra-cluster correlation is observed, i.e. for cases CSNI3 and CSNI4 when $\delta = 1$; but it still remains clear in the tables and the marginal estimator, which ignore clustering, is the worst of all estimators under cluster-specific nonignorable nonresponse.

Regarding the random effect estimator, there is evidence of negative bias of $\hat{T}_y^{RE}$ under MCAR and MAR ($\delta = 0$) and also under CSNI3 and CSNI4 ($\delta = 1$) where the relative bias of the random effect estimator moves in the direction towards its bias when $\delta = 0$ (the MCAR and MAR cases). In the cluster-specific nonignorable cases with higher intra-cluster correlation (CSNI1 and CSNI2 when $\delta = 5$), $\hat{T}_y^{RE}$ displays bias in the same positive direction as $\hat{T}_y^{M}$.

Table 3.5.1: Simulation estimates of relative bias, standard errors and root mean squared errors of weighted estimates of totals for alternative weighting methods and missing data mechanisms. Cluster sampling with $n = 50$, $m_i = 10$. Simulation estimates based on 1000 repeated samples.

Missing data mechanism	Weighting Method	Relative Bias (%)[1]	Relative SE (%)	Relative RMSE (%)
MCAR	Response prop. (M)	(-0.13)	2.33	2.33
	Random effects (RE)	-2.76	2.35	3.63
	Fixed effects (FE)	(0.04)	2.48	2.48
	CML (estimated)	(0.04)	2.48	2.48
	CML (true parameter)	(0.01)	2.92	2.92
MAR	Response prop. (M)	(-0.14)	2.34	2.34
	Random effects (RE)	-2.35	2.32	3.30
	Fixed effects (FE)	(0.07)	2.34	2.34
	CML (estimated)	(0.07)	2.34	2.34
	CML (true parameter)	(0.08)	2.48	2.48
CSNI1	Response prop. (M)	11.10	6.19	12.71
	Random effects (RE)	2.20	6.05	6.44
	Fixed effects (FE)	(-0.14)	6.25	6.25
	CML (estimated)	(-0.14)	6.25	6.25
	CML (true parameter)	(-0.19)	6.43	6.43
CSNI2	Response prop. (M)	11.35	6.19	12.92
	Random effects (RE)	2.67	6.00	6.57
	Fixed effects (FE)	(-0.13)	6.33	6.33
	CML (estimated)	(-0.12)	6.32	6.32
	CML (true parameter)	(-0.13)	6.43	6.43
CSNI3	Response prop. (M)	2.19	2.50	3.32
	Random effects (RE)	-1.74	2.61	3.13
	Fixed effects (FE)	(0.01)	2.68	2.68
	CML (estimated)	(0.01)	2.68	2.68
	CML (true parameter)	(-0.03)	3.09	3.09
CSNI4	Response prop. (M)	2.23	2.50	3.35
	Random effects (RE)	-1.31	2.56	2.88
	Fixed effects (FE)	(0.03)	2.57	2.57
	CML (estimated)	(0.03)	2.57	2.57
	CML (true parameter)	(0.03)	2.72	2.72

[1] parentheses surround estimates which are within two simulation standard errors of 0.

Table 3.5.2: Simulation estimates of relative bias, standard errors and root mean squared errors of weighted estimates of totals for alternative weighting methods and missing data mechanisms. Two-stage sampling with $n = 50$, $m_i = 5$. Estimates based on 1000 repeated samples.

Missing data mechanism	Weighting Method	Relative Bias (%)[1]	Relative SE (%)	Relative RMSE (%)
MCAR	Response prop. (M)	(-0.06)	3.20	3.20
	Random effects (RE)	-3.08	3.33	4.54
	Fixed effects (FE)	(0.18)	3.65	3.66
	CML (estimated)	(0.17)	3.56	3.57
	CML (true parameter)	(0.13)	3.95	3.96
MAR	Response prop. (M)	(-0.02)	3.13	3.13
	Random effects (RE)	-2.57	3.20	4.11
	Fixed effects (FE)	(0.19)	3.23	3.24
	CML (estimated)	(0.20)	3.24	3.24
	CML (true parameter)	(0.14)	3.39	3.40
CSNI1	Response prop. (M)	10.39	6.56	12.29
	Random effects (RE)	3.55	6.44	7.36
	Fixed effects (FE)	(-0.04)	6.66	6.66
	CML (estimated)	(-0.04)	6.64	6.64
	CML (true parameter)	(-0.06)	7.02	7.02
CSNI2	Response prop. (M)	10.75	6.61	12.62
	Random effects (RE)	4.21	6.37	7.64
	Fixed effects (FE)	(-0.05)	6.77	6.77
	CML (estimated)	(-0.01)	6.74	6.74
	CML (true parameter)	(-0.10)	6.97	6.97
CSNI3	Response prop. (M)	2.10	3.31	3.92
	Random effects (RE)	-1.71	3.49	3.89
	Fixed effects (FE)	(0.14)	3.73	3.73
	CML (estimated)	(0.13)	3.65	3.66
	CML (true parameter)	(0.09)	4.10	4.10
CSNI4	Response prop. (M)	2.21	3.25	3.93
	Random effects (RE)	-1.17	3.37	3.57
	Fixed effects (FE)	(0.14)	3.40	3.40
	CML (estimated)	(0.16)	3.40	3.40
	CML (true parameter)	(0.09)	3.61	3.61

[1] parentheses surround estimates which are within two simulation standard errors of 0.

The bias of the random effect estimator under cluster-specific non-ignorable nonresponse is always of smaller size than this of the marginal estimator, in particular for the CSNI1 and CSNI2 cases. The simulation study shows that the random effect estimator leads to some bias even for the missing at random mechanism when m_i are small. A theoretical explanation in the bias of the RE method is available in Skinner and D'Arrigo (2011, Section 4) and relates to the potential correlation between the response indicator variable R_{ij} and the estimated random effect $\hat{u}_i^{RE}$. Problems with the

approximate maximum likelihood estimation method used to obtain $\hat{u}_i^{RE}$ may be another source of the observed bias. Table 3.5.2 and 3.5.1 shows a small decline in the relative bias of $\hat{T}_y^{RE}$ under MAR when m_i increases from 5 (-2.57) to 10 (-2.35) respectively. Repeating this study for $m_i = 20$ and $m_i = 50$, it is observed that the relative bias of $\hat{T}_y^{RE}$ does indeed decrease as m_i increases, with values of -1.67 and -0.78 as m_i takes values 20 and 50 correspondingly. The empirical illustration in section 3.6 examines the performance of the random effect estimator using data from the Expenditure and Food Survey 2001 with cluster mean size equals 40.

In line with the theoretical results presented by Skinner and D'Arrigo (2011), Table 3.5.1 and 3.5.2 show no evidence of bias in $\tilde{T}_y^{CML}$ or $\hat{T}_y^{CML}$ across all missing data mechanisms and sampling schemes. However, one potential disadvantage to this conditional maximum likelihood approach is that it becomes increasingly computationally intensive for larger m_i as the sizes of the sets B_{1ij} and B_{2i} in (3.3.1) grow.

Regarding the fixed effects estimator, it is observed in Table 3.5.1 and 3.5.2 that $\hat{T}_y^{FE}$ seems to share a similar absence of bias to $\hat{T}_y^{CML}$, which may be attractive in practice as this estimator does not require so much computation.

Looking now at the standard errors of the weighted estimates of total, there is some evidence in Table 3.5.1 and 3.5.2 that the variance of $\hat{T}_y^{RE}$ and $\hat{T}_y^{M}$ can be slightly smaller than those of $\hat{T}_y^{CML}$ and $\hat{T}_y^{FE}$ for all cases. However, the smaller biases of the conditional ML and the fixed effect estimators offset this effect. The RMSEs of the latter estimators are always smaller than that of $\hat{T}_y^{RE}$ and they are also considerably smaller than that of $\hat{T}_y^{M}$ for the CSNI cases. The extent to which the smaller bias of $\hat{T}_y^{CML}$ will offset its larger variance, in MSE terms, will, of course, depend on sample size. The RMSE of $\hat{T}_y^{RE}$ is smaller than that of $\hat{T}_y^{M}$ for all cluster-specific nonignorable cases, in particular substantially smaller for CSNI cases with higher intra-cluster correlation (CSNI1 and CSNI2).

Comparing $\hat{T}_y^{CML}$ and $\tilde{T}_y^{CML}$, there are some results in the literature (e.g. Rosenbaum, 1987; Kim and Kim, 2007) that the use of the estimated rather than the

true response propensity can, paradoxically, reduce variance. This is observed in Tables 3.5.1 and 3.5.2 where the relative standard error of $\hat{T}_y^{CML}$ is smaller than that of $\tilde{T}_y^{CML}$ for all cases.

Under cluster-specific nonignorable nonresponse, this simulation study shows some potential benefits of multilevel or fixed effects models or conditional logistic regression over marginal models (which ignore clustering). Benefits are greater the larger the cluster sample size. On the other hand, this simulation shows no benefits from using methods that account for the clustering in the data under the missing at random assumption.

Generalized regression point estimate

Table 3.5.3 shows results on the regression estimator $\hat{T}_{yreg}$ in 3.2.2. Results for $\hat{T}_{yreg}^{M}$ were almost identical to those for $\hat{T}_y^M$, and are thus not included in the table. Results for $\hat{T}_{yreg}^{FE}$ and $\tilde{T}_{yreg}^{CML}$ were almost identical to those for $\hat{T}_{yreg}^{CML}$ and are also thus not included, although it is of interest to note that the reduction in variance of $\hat{T}_{yreg}^{CML}$ vs. $\tilde{T}_{yreg}^{CML}$ observed in Tables 3.5.1 and 3.5.2 seems to disappear once regression estimation is used.

Table 3.5.3 shows that the bias of the multilevel estimator under MCAR and MAR is removed by regression estimation. However, it remains biased under the CSNI mechanisms with larger values for cases when $\delta = 5$ and smaller ones when $\delta = 1$. As expected, regression estimation does lead to some overall reduction in variance as it borrows strength from a linear assisting model. As in previous tables, $\hat{T}_{yreg}^{RE}$ does show some slight variance gains relative to $\hat{T}_{yreg}^{CML}$ but this is offset by bias and the RMSE of $\hat{T}_{yreg}^{CML}$ is in no cases greater than that of $\tilde{T}_{yreg}^{RE}$.

Table 3.5.3: Simulation estimates of relative bias, standard errors and root mean squared errors of regression weighted estimates of totals for alternative weighting methods and missing data mechanisms. Estimates based on 1000 repeated samples.

Missing Data mechanism	Weighting Method	Relative Bias (%)[1]	Relative SE (%)	Relative RMSE (%)
$n = 50, m_i = 10$				
MCAR	Random effects (RE)	(-0.02)	2.29	2.29
	CML (estimated)	(0.06)	2.29	2.29
MAR	Random effects (RE)	(-0.01)	2.29	2.29
	CML (estimated)	(0.07)	2.30	2.30
CSNI1	Random effects (RE)	5.02	5.94	7.77
	CML (estimated)	(-0.13)	6.18	6.18
CSNI2	Random effects (RE)	5.06	5.90	7.78
	CML (estimated)	(-0.12)	6.28	6.28
CSNI3	Random effects (RE)	1.02	2.47	2.67
	CML (estimated)	(0.02)	2.51	2.51
CSNI4	Random effects (RE)	1.04	2.46	2.67
	CML (estimated)	(0.03)	2.52	2.52
$n = 50, m_i = 5$				
MCAR	Random effects (RE)	(0.00)	3.18	3.18
	CML (estimated)	(0.11)	3.17	3.17
MAR	Random effects (RE)	(0.07)	3.10	3.10
	CML (estimated)	(0.19)	3.12	3.12
CSNI1	Random effects (RE)	6.77	6.29	9.24
	CML (estimated)	(-0.10)	6.50	6.50
CSNI2	Random effects (RE)	6.95	6.25	9.35
	CML (estimated)	(-0.03)	6.66	6.66
CSNI3	Random effects (RE)	1.40	3.28	3.57
	CML (estimated)	(0.07)	3.30	3.30
CSNI4	Random effects (RE)	1.49	3.21	3.54
	CML (estimated)	(0.14)	3.28	3.28

[1] parentheses surround estimates which are within two simulation standard errors of 0.

Variance estimation of the alternative adjusted estimates

Table 3.5.4 presents results on the estimation of the variance of the alternative weighted estimators for the case of cluster sampling and treating the weights $\hat{q}_{ij}$ as fixed. The variance estimator in (3.4.1) is used for each estimator under study, including a finite population correction term $(1 - n/N)$.

The variance estimate for the conditional maximum likelihood estimator and the fixed effect estimator are always of considerably smaller size than those for the other estimators, with the variance of the marginal estimator performing the worst. Skinner and D'Arrigo (2011) showed that allowing for variation in $\hat{\beta}$ reduces the variance estimate of $\hat{T}_y^{CML}$; however, their variance estimator is more complex to compute and result in some underestimation, if modest, under the MCAR and MAR missing mechanism. In some applications, it may be attractive to obtain simpler variance estimators that are always conservative.

It might be desirable to consider an alternative variance estimate for the random and fixed effect estimators that allows for variation in the weights, such as jackknife variance estimation as described in Chapter 4.

Table 3.5.4: Simulation estimates of relative bias, standard errors and root mean squared errors of standard error estimators for alternative weighting estimation of totals (treating weights as fixed) and missing data mechanisms. Cluster sampling with $n = 50$, $m_i = 10$. Simulation estimates based on 1000 repeated samples.

Missing data mechanism	Weighting Method	Relative Bias (%)[1]	Relative SE (%)	Relative RMSE (%)
MCAR	Response prop. (M)	92.41	19.42	94.43
	Random effects (RE)	30.03	13.10	32.77
	Fixed effects (FE)	10.70	13.76	17.43
	CML (estimated)	10.71	13.75	17.43
MAR	Response prop. (M)	75.64	17.68	77.68
	Random effects (RE)	22.49	11.72	25.36
	Fixed effects (FE)	3.93	10.29	11.01
	CML (estimated)	3.88	10.26	10.97
CSNI1	Response prop. (M)	49.31	14.07	51.28
	Random effects (RE)	20.14	9.46	22.25
	Fixed effects (FE)	3.05	10.34	10.78
	CML (estimated)	3.06	10.34	10.78
CSNI2	Response prop. (M)	44.66	14.02	46.81
	Random effects (RE)	18.12	9.64	20.52
	Fixed effects (FE)	1.38	13.27	13.34
	CML (estimated)	1.44	13.06	13.14
CSNI3	Response prop. (M)	108.46	19.31	110.16
	Random effects (RE)	37.51	11.88	39.35
	Fixed effects (FE)	9.44	13.15	16.18
	CML (estimated)	9.44	13.14	16.18
CSNI4	Response prop. (M)	93.14	18.01	94.86
	Random effects (RE)	31.06	11.10	32.99
	Fixed effects (FE)	3.76	12.47	13.02
	CML (estimated)	3.76	12.36	12.92

[1] parentheses surround estimates which are within two simulation standard errors of 0.

3.6 Empirical application

This section presents an application of the various estimators presented in the previous sections to data from the Expenditure and Food Survey (EFS) collected in 2001. It aims to estimate three parameters: the proportion of households with at least one adult in employment, the proportion of households with at least one pensioner and the proportion of single households in the UK using inverse-probability weighted estimators with alternative 'standard' inverse weights, that is response propensity weights (M), weights based on predicted random effects (RE) and weights based on estimated fixed effects (FE). The CML approach is not considered in this empirical illustration as it becomes extremely computationally intensive with large cluster sizes. The EFS, which is part of the UK 2001 Census Link Study dataset presented in Chapter 2, provides information on the pattern of spending and food consumption by households in the UK. In addition to expenditure and food intake, the EFS collects socio-demographic information about households, such as household composition and employment details. The EFS employs a multi-stage stratified random sampling design and requires a face-to-face interview and the filling in of a diary. As described in Chapter 2, the response outcome of the EFS data from April to October 2001 was linked to the 2001 Census records, which are available for both respondents and nonrespondents to the EFS, providing a rare opportunity to model nonresponse and in turn to adjust for it.

The analysis sample for this illustration includes 2994 households selected for interviewing in the EFS and for which the survey outcome was successfully linked to census information and interviewer observation data and the interviewer could be identified. Cases such as vacant homes and reissues as well as cases were the survey outcome could not be linked to census or interviewer information have been deleted. The actual survey variables were not included by the ONS in the dataset. Thus, census variables are used in this application as if they were measured in the survey. The unit nonresponse rate, which this section aims to adjust for, is about 35%. The estimates in this application do not attempt to adjust for the complex sampling scheme as sampling weights are not available in the dataset. The clusters are defined by interviewer workloads, with 130 clusters of mean size 23 households. Each cluster contains at least 10 and at most 49 households.

To obtain inverse estimated response probabilities for weighting adjustment purposes, this application first models the response indicator R_{ij}, with refusals and

noncontacts both coded as nonresponse, considering the three 'standard' approaches described in previous sections: (1) a marginal model (3.2.4); (2) a random effect model (3.2.3); and (3) a fixed effects model treating cluster as another explanatory variable. The inverse logit function is used as the inverse link function for all models. Table 3.6.1 presents estimated coefficients and standard errors under the three models for nonresponse. Model 3 also produces estimated coefficients for each cluster, but these are not presented here for space reasons.

Table 3.6.1: Estimated coefficients (and standard errors) of the three logistic models modelling response

Variable (ref= Reference category)	**Categories**	**Model 1 (M)** $\hat{\beta}$ $(ste(\hat{\beta}))$	**Model 2 (RE)** $\hat{\beta}$ $(ste(\hat{\beta}))$	**Model 3 (FE)** $\hat{\beta}$ $(ste(\hat{\beta}))$
Constant		0.664 (0.167)***	0.697 (0.173)***	0.636 (0.421)
Interviewer Observations				
Type of accommodation (ref= Not house, i.e. flat, mobile home, other)	House	0.272 (0.110)***	0.237 (0.112)**	0.200 (0.121)*
House in a better or worse condition than others in area (ref= Better)	About the same Worse	-0.263 (0.136)** -0.700 (0.193)***	-0.269 (0.139)** -0.740 (0.196)***	-0.287 (0.148)** -0.876 (0.209)***
Household-level variables from the Census				
Dependent children present (ref= Not present)	Present	0.442 (0.089)***	0.454 (0.090)***	0.486 (0.095)***
London indicator (ref =Not London)	London	-0.512 (0.128)***	-0.508 (0.160)***	-0.374 (0.433)
Self-employment indicator of HRP (ref = Not self-employed)	Self-employed	-0.552 (0.134)***	-0.564 (0.135)***	-0.631 (0.143)***
Educational attainment of HRP (ref=A levels/GCSEs)	First/Higher degree No academic qualifications	0.407 (0.112)*** -0.215 (0.089)**	0.423 (0.114)*** -0.205 (0.090)**	0.492 (0.120)*** -0.219 (0.096)**
Interviewer variance		---	0.107 (0.038)**	---

HRP household representative person
* significant at the 10% level
** significant at the 5% level
*** significant at the 1% level

Table 3.6.1 shows that, regardless of the nonresponse model, there are several factors significantly influencing nonresponse. This indicates that nonresponse is not MCAR. This table displays similar estimated fixed coefficients under the three models. Model 2 shows a significant between-interviewer (between-cluster) variance, indicating that nonresponse depends on unobserved interviewer effects. It is important to note that nonresponse depending on cluster effects is not necessarily non-ignorable. It will only be non-ignorable if the cluster effects are correlated with the survey variable of

interest, meaning that the MAR assumption fails. Model 2 allows the calculation of the variance partitioning coefficient or intra-cluster correlation, which indicates the percentage of observed variation in response attributable to interviewer characteristics. The intra-cluster correlation coefficient for Model 2, using the idea of the threshold model for logit model (Snijders and Bosker, 1999), is equal to 0.107/(3.29+0.107)= 0.031. This variance partitioning coefficient indicates that about 3% of the variance on response rates is attributable to interviewer characteristics. Comparing with Model 1, Model 3 shows that controlling for cluster fixed effects some demographic variables, such as London or house indicator, become not significant at the 5% level.

For any of the variables in Table 3.6.1 to be related to nonresponse bias they have to be associated with both nonresponse and the survey variable. Table 3.6.2 presents the estimated coefficients obtained by regressing the first survey variable of interest, 'households with at least one adult in employment', on the above explanatory variables. Note that the variable 'self-employment of HRP' included in the response model (Table 3.6.1) is not included in Table 3.6.2 due to high correlation with the outcome variable of interest.

Table 3.6.2: Estimated coefficients (and standard errors) for three logistic models for the indicator household with at least one adult in employment

Variable (ref= Reference category)	**Categories**	**Model 4 (M)** $\hat{\beta}$ $(ste(\hat{\beta}))$	**Model 5 (RE)** $\hat{\beta}$ $(ste(\hat{\beta}))$	**Model 6 (FE)** $\hat{\beta}$ $(ste(\hat{\beta}))$
Constant		-0.423 (0.169)***	-0.443 (0.172)***	-1.150 (0.445)***
Interviewer Observations				
Type of accommodation (ref= Not house, i.e. flat, mobile home, other)	House	0.588 (0.114)***	0.599 (0.115)***	0.689 (0.125) ***
House in a better or worse condition than others in area (ref= Better)	About the same Worse	-0.076 (0.135) 0.029 (0.201)	-0.065 (0.137) 0.037 (0.203)	-0.033 (0.149) 0.057 (0.216)
Household-level variables from the Census				
Dependent children present (ref= Not present)	Present	1.414 (0.097)***	1.414 (0.098)***	1.494 (0.104)***
London indicator (ref =Not London)	London	-0.105 (0.137)	-0.108 (0.157)	-0.852 (0.462)*
Educational attainment of HRP (ref=A levels/GCSEs)	First/Higher degree No academic qualifications	1.022 (0.123)*** -0.645 (0.090)***	1.016 (0.123)*** -0.639 (0.091)***	1.041 (0.130)*** -0.652 (0.097) ***
Interviewer variance		---	0.064 (0.033)**	---

HRP household representative person
* significant at the 10% level
** significant at the 5% level
*** significant at the 1% level

Table 3.6.2 shows that some of the variables significant on the response model are also significant to predict the outcome variable of interest. In addition, Model 5 indicates that the survey variable of interest also depends on unobserved cluster random effects. Similar results are found for the other two survey variables of interest (results not shown). The combination of findings from Tables 3.6.1 and 3.6.2 imply that the unweighted estimator may be subject to some bias.

To further assess the response mechanism, the estimated interviewer random effects from Model 5, which measure the difference between the average number of households with at least one adult in employment reported by an interviewer and the average number of households with at least one adult in employment in the whole sample, are plotted against the estimated interviewer random effects from the nonresponse Model 2 (Fig 3.6.1). This scatterplot seems to show no systematic pattern or clear correlation between the random effects from the two models. Therefore, it suggests that, provided the mechanism is CSNI, it would appear that a MAR assumption is reasonable. The fact that this correlation is not evident means that the M or RE estimators may be unbiased in this application.

Figure 3.6.1: Estimated random effects from model for survey variable against estimated random effects from nonresponse model

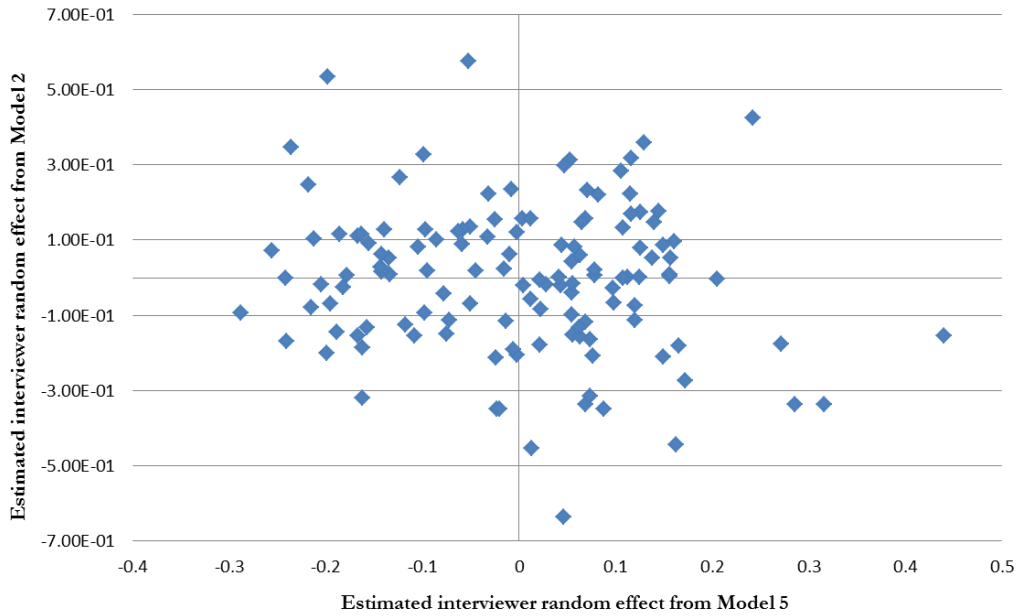


Table 3.6.3 present the estimates of the three parameters of interest (proportion of households with at least one adult in employment, the proportion of households with

at least one pensioner and the proportion of single households in the UK) for the three weighting methods. The table also presents the unweighted estimates based on respondents only and the true proportion derived for the whole sample, i.e. based on respondents and nonrespondents to the survey. The latter is only possible because of the availability of the Census data.

Table 3.6.3 shows that nonresponse approximately accounts for a 1.5% point difference in the employment estimate (60.20% vs. 58.75%), about 2% point difference in the pensioners estimate (31.07% vs. 33.03%), and a 1.7% point difference in the single households estimate (28.43% vs. 30.16%). All weighting methods yield similar results across different estimates. In particular, the RE method does not appear seriously biased in this example compared to the other approaches. The results do not indicate any consistent gains for the RE or FE approach compared to weighting using the marginal model (i.e. ignoring clustering) in line with the theory that suggests that under the MAR assumption there is little to be gained from the method that account for the clustering in the data over the M approach (Skinner and D'Arrigo, 2011). Still Table 3.6.3 also does not show any disadvantages from the RE method compared to the M method as observed in the simulation work under the MAR nonresponse mechanisms. It should be noted, however, that the conditions here in this application are somewhat different than those in the simulation study, for example, in this application the clusters are of unequal large sizes.

Standard errors to the estimates are not included in Table 3.6.3 due to the fact that the simulation work shows severe bias on the variance estimates that treat weights as fixed, in particular for M and RE. Therefore, this application cannot assess how far the differences might be attributable to sampling variation.

Table 3.6.3: Estimates of proportion of households with at least one adult in employment, proportion of households with at least one pensioner and proportion of single households by various weighting methods using data from the EFS

	Estimate		
	Employment %	Pensioners %	Single %
True value based on the whole sample (respondents and nonrespondents)	58.75	33.03	30.16
Unweighted estimator	60.20	31.07	28.43
Weighted: Response prop. (M)	58.21	32.68	30.10
Weighted: Random effects (RE)	58.43	32.61	30.03
Weighted: Fixed effects (FE)	58.63	32.54	30.11

3.7 Conclusions

This chapter proposes different ways of constructing inverse probability weights to estimate the finite population total under clustered nonresponse. It compares the properties of the alternative weighted estimators for two sampling designs by a simulation study and presents results from an empirical application using data from the Expenditure and Food Survey 2001.

The simulation study shows that an effort to allow for clustered response via the introduction of predicted random effects into the estimated probability of response can actually induce negative relative bias in the inverse probability weighted estimator under MAR, when the cluster sizes are not large. For example, a relative bias of about 2% for the random effect estimator for small cluster sizes of between 5 and 20 is observed in the simulation study. This bias declines to about 1% as the cluster sizes increased to 50. Although the empirical application does not show any disadvantages from the random effect method compared to the other methods, it does not show any advantage either. Therefore, if MAR is plausible, it seems reasonable to employ simple response propensity weights based upon a marginal model for response rather than weights based on a multilevel model.

If nonresponse is CSNI but not MAR then the marginal approach may be subject to bias, in particular higher relative biases of about 11% are observed when allowing for high intra-cluster correlation in both the survey variable and the nonresponse process. The proposed CML approach performs the best and removes this bias, when the number of sampled clusters is large even if the cluster sizes are small. In the simulation study it is also observed that the fixed effects estimator performed similarly to the CML estimator and it may be that in practice it will often provide a reasonable proxy to this estimator, while not requiring such strong model assumptions nor so much computation. Regarding the use of multilevel models to construct inverse probability weights under CSNI, the simulation results show some potential benefits of the random effects estimator over the simple response propensity estimator based on a marginal model, in particular for larger cluster sample sizes.

In addition to its bias correction advantage, the CML approach is not dependent on the assumption that the u_i term in (3.2.3) is Gaussian, nor that it is independent of x_{ij}. There are, however, potential disadvantages to the CML approach. It depends on

the logistic form of the model in (3.2.3) and becomes increasingly computationally intensive as the sizes of the sets B_{1ij} and B_{2i} grow. In addition, as observed in the simulation study, it can lead to more variable weights and have efficiency disadvantages.

Regarding efficiency, the simulation study shows that the simple variance estimator for the conditional maximum likelihood estimate, treating weights as fixed, is always conservative and of considerably smaller size than those for the other estimators. The variance of the marginal and random effect estimators perform the worst. For these cases, it would be advisable to consider a variance estimator that account for the nonresponse adjustments.

Chapter 4

Variance Estimation for Calibration Weighted Estimators in the Presence of Nonresponse

4.1 Introduction

Weighting methods that make use of auxiliary information are widely used to compensate for potential bias caused by survey nonresponse. However, a concern with these methods is that they might result in increased variability in the weights and thereby lower the precision of the survey estimates. Therefore, it is necessary to consider the effects on bias and variance of the estimates resulting from using different weighting adjustments when comparing their relative properties. This chapter focuses on a particular type of weighting procedure called calibration. Deville et al. (1993) proposed a class of calibration methods, called generalized raking estimation, which can be used for estimation in surveys with auxiliary information in the form of known population totals. The generalized raking weights have the property to reproduce the known population totals when applied to each auxiliary variable. Therefore, a strong correlation between the auxiliary variables and the survey variable is essential for the weights to perform efficiently on the study variable too. The auxiliary information used for weighting may come from one or more external sources, such as administrative data files or census data. In some surveys there is also information at the sample level (i.e. for both respondents and nonrespondents) on auxiliary variables. For simplicity, this chapter will assume auxiliary information as a set of variables that have been measured on respondents to the survey and for which information on the population totals is available.

In this chapter three forms of generalized raking estimator in the presence of nonresponse are discussed: the generalized regression estimator (GREG), the classical raking ratio estimator and the 'maximum likelihood' raking estimator (Brackstone and Rao, 1979; Fuller, 2002). These estimators are designed to take account of differences in the characteristics of respondents on a set of auxiliary variables with the characteristics of the population. Deville and Särndal (1992) and Deville et al. (1993), showed that,

under their full response setting and framework, the GREG estimator and the classical raking estimator have asymptotically the same properties.

Raking estimation appears to have a more well-established history of applications in many national statistical institutes (NSIs), perhaps because of its ease of computation, involving repeated use of standard post-stratification adjustments (Kalton and Flores-Cervantes, 2003). In some NSIs, GREG has tended to replace raking estimation, and is now used in many surveys (Särndal and Lundström, 2005). One reason is that the GREG can be expressed in closed form and computed in one step, whereas the computation of a raking estimator is iterative. Perhaps a more important reason is that GREG can handle a wider class of auxiliary information, including population totals of continuous variables, whereas raking is restricted to the use of population counts in the categories of discrete variables. Nevertheless, raking estimation continues to be widely used in NSIs in many countries, e.g. the USA and the UK. One advantage is that it always produces positive weights, whereas GREG requires modification to meet this condition. In addition, raking may reduce nonresponse bias more than GREG under certain assumptions (Kalton and Flores-Cervantes, 2003).

The variances of weighted estimators are often estimated using linearization methods (Demnati and Rao, 2004; Wolter, 2007), which rely on the validity of Taylor series expansions, or replication techniques (Efron, 1981; Wolter, 2007), which treat the sample as it if were the population and repeatedly subsample from this population to estimate a variance. A simulation study by Stukel et al. (1996) found little difference between two forms of linearization estimators with respect to sampling and observed that both the linearization and the jackknife variance estimators show small underestimation of the true variance. Stukel et al. (1996) also noted that the jackknife approach consistently had smaller biases than the linearization one. However, Stukel et al. (1996) simulation work was designed for the full response set-up and there are reasons why in the presence of nonresponse different results may be expected. A simulation study by Valliant (2004) observed negatively biased linearization variance estimators contrary to positively bias jackknife replication variance estimators.

Conditions for unbiasedness of raking estimation methods under nonresponse models vary between estimation methods (e.g. Kalton and Maligalig, 1991; Kalton and Flores-Cervantes, 2003) and the choice of variance estimators may be more important in the presence of nonresponse (e.g. Fuller, 2002, Sect.8).

This chapter explores alternative forms of linearization variance estimators for generalized raking estimators in the presence of unit nonresponse. It also investigates one of the most frequently used replication methods of computing variances for complex sample surveys called the jackknife method.

The properties of the alternative raking estimators, including bias and root mean square error, and associated variance estimators are investigated through a simulation study. This study is designed to mimic these properties with respect to the effects of both sampling and nonresponse for two European surveys conducted by NSIs: the British Labour Force Survey (LFS) and the German Survey of Income and Expenditure (SIE). The GREG estimator is used in practice in the LFS while a version of the 'maximum likelihood' raking estimator is employed in the SIE.

The chapter is structured as follows. Section 4.2 defines the generalized raking estimators. Linearization variance estimators are defined in section 4.3 and replication variance estimators in section 4.4. Section 4.5 presents the simulation study with results discussed in section 4.6. Some concluding remarks are given in section 4.7.

4.2 Generalized raking estimation

Let us first define the three forms of generalized raking estimator in the presence of unit nonresponse: a) generalized regression estimator (GREG), b) classical raking ratio estimator and c) 'maximum likelihood' raking estimator. Consider a finite population U from which a probability sample s is drawn with a given sample design. However, as nonresponse occurs, the response set r is obtained, where $r \subseteq s$. The objective is to estimate the population total $T_y = \sum_{i \in U} y_i$, where y_i is the value of a survey variable for the ith population element, with which is also associated an auxiliary vector value $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ})'$. The population total of $\boldsymbol{x}$, $\boldsymbol{T}_x = \sum_{i \in U} \boldsymbol{x}_i$, is supposed to be accurately known and $\boldsymbol{x}_i$ is known for all units in r. Following Deville and Särndal (1992) and Deville et al. (1993), a generalized raking estimator for the population total T_y may be written as

$$\hat{T}_y = \sum_{i \in r} w_i y_i , \tag{4.2.1}$$

where the *calibration weight* w_i are as close as possible, according to a specified distance function, to the *initial weights* d_i while satisfying the *calibration equation*:

$$\sum_{i \in r} w_i \boldsymbol{x}_i = \boldsymbol{T}_x . \tag{4.2.2}$$

The vector $\boldsymbol{T}_x$ is referred to as the vector of *calibration totals*. A common choice of initial weights, which is taken here, is the design weights, i.e. $d_i = \pi_i^{-1}$, where π_i is the probability that unit i is sampled.

Let $G(.)$ be the distance function from the calibrated weight w_i to the initial weight d_i, with argument w_i / d_i. For every fixed $d_i > 0$, it is assumed that $G(.)$ is positive, differentiable with respect to w_i, strictly convex, $G\ 1\ = G'\ 1\ = 0$, implying that when $w_i = d_i$ the distance between the weights is zero, and $G''\ 1\ > 0$, which makes $w_i = d_i$ a local minimum (Deville and Särndal, 1992; Deville et al., 1993). The class of generalized raking weights w_i is obtained by minimising the total sample distance

$$\sum_{i \in r} d_i G(w_i / d_i), \tag{4.2.3}$$

subject to the calibration equation (4.2.2). Explicitly, if λ denotes a vector of Lagrange multipliers, the expression

$$\sum_{i \in r} d_i G(w_i / d_i) - \lambda' \left(\sum_{i \in r} w_i \boldsymbol{x}_i - \boldsymbol{T}_x \right) \tag{4.2.4}$$

is minimized with respect to the w_i. Differentiating (4.2.4) with respect to w_i and equating to zero results $g(w_i / d_i) - \boldsymbol{x}_i' \lambda = 0$, where $g\ u\ = dG\ u\ / du$, and solving for w_i leads to the calibration weights:

$$w_i = d_i F(\boldsymbol{x}_i' \hat{\lambda}), \tag{4.2.5}$$

where $F(u) = g^{-1}(u)$ denotes the inverse function of $g\ u$ and $\hat{\lambda}$ is the Lagrange multiplier which solves the calibration equations:

$$\sum_r d_i\ F\ (\boldsymbol{x}_i' \hat{\lambda}) \boldsymbol{x}_i = \boldsymbol{T}_x .$$

Various choices of the distance function $G(.)$ and associated function $F(.)$ are discussed by Deville and Särndal (1992) (see also Deville et al., 1993 and Fuller, 2009, Sect. 2.9). Three are considered in this chapter, which lead to the different generalised raking estimators, as follows:

a) linear: $G_L(u) = (1/2)(u-1)^2,\ F_L(u) = 1+u$

b) multiplicative (raking ratio): $G_M\ u\ = u \log\ u\ -\ u+1,\ F_M\ u\ = \exp\ u$

c) maximum likelihood raking: $G_{ML}\ u\ =\ u-1-\log\ u\ ,\ F_{ML}\ u\ =\ 1-u^{-1}$

Minimization of expression (4.2.4) using the linear choice of $G(.)$, $G_L(u)$, leads to the calibration weights:

$$w_i = d_i\left[1+\ \boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{xd}}\ '\left(\sum_{i\in r} d_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \boldsymbol{x}_i\right], \tag{4.2.6}$$

where $\hat{\boldsymbol{T}}_{xd} = \sum_{i\in r} d_i \boldsymbol{x}_i$, and the generalized raking estimator becomes

$$\hat{T}_y = \sum_{i\in r} w_i y_i = \hat{T}_{yd} + (\boldsymbol{T}_x - \hat{\boldsymbol{T}}_{xd})'\hat{\boldsymbol{B}}_r, \tag{4.2.7}$$

the *generalised regression estimator* (GREG), where $\hat{\boldsymbol{T}}_{yd} = \sum_{i\in r} d_i y_i$ and

$$\hat{\boldsymbol{B}}_r = \left(\sum_{i\in r} d_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_{i\in r} d_i \boldsymbol{x}_i y_i\,. \tag{4.2.8}$$

With the second option, the multiplicative choice of $G(.)$, $G_M\ u\ $, the calibrated estimator of T_y is the *classical raking ratio estimator* (Brackstone and Rao, 1979) when T_x contains the population counts in the categories of two or more categorical auxiliary variables. For example, in the context of the British Labour Force Survey, $\boldsymbol{x}_i$ denotes the vector of indicator variables of three categorical auxiliary variables:

$$\boldsymbol{x}_i = (\delta_{1..i},\ \ldots,\ \delta_{A..i},\ \delta_{.1.i},\ \ldots,\ \delta_{.Bi},\ \delta_{..1i},\ \ldots,\ \delta_{..Ci})',$$

where $\delta_{a..i} = 1$ if unit i is in category a of the first auxiliary variable and 0 otherwise, $\delta_{b.i} = 1$ if unit i is in category b of the second auxiliary variable and 0 otherwise and so on. The population total T_x of this vector thus contains the population counts in each of the (marginal) categories of each of the three auxiliary variables. The construction of

the weights for classical raking ratio estimation has traditionally involved the use of iterative proportional fitting (Brackstone and Rao, 1979). Ireland and Kullback (1968) demonstrate that this method converges to a solution of the above optimisation problem.

The third option, the function $G_{ML}\ u$, leads to an alternative 'maximum likelihood' version of raking adjustment, when $\boldsymbol{x}_i$ takes the same form, denoting indicator variables of categorical auxiliary variables. In this case, the distance (4.2.3) may be interpreted as a quantity which is proportional to minus a log likelihood in the case of simple random sampling with replacement (Brackstone and Rao, 1979; Fuller, 2002).

A disadvantage of using a linear form of $G(.)$ compared to the other choices presented in this section is that, as noted by Deville and Särndal (1992), the calibrated weights w_i resulting from using the linear function can be positive or negative, whereas the multiplicative and 'maximum likelihood' cases guarantee positive weights. Deville and Särndal (1992) also noted that the multiplicative choice of $G(.)$ may result in some extremely large weights compared to the basic sampling weights d_i.

This chapter now turns to the discussion of variance estimation methods for the generalised raking estimators, including both linearization and replication variance estimation.

4.3 Linearization variance estimation

Survey weights that include calibration for nonresponse should not be treated as constants when estimating the variances of survey estimates since they are sample dependent. One possible approach to deal with this complication is to use linearization variance estimators (Wolter, 2007). This approach is usually called the linearization method because one first reduces the original nonlinear quantity to an approximate linear quantity by using the linear term of the corresponding Taylor series expansion, and then constructs the variance formula and an estimator of the variance of this linearized quantity.

Suppose first that $\hat{\theta}$, an estimator of a population parameter θ based on a sample s of size n, may be expressed as a linear function of p estimated totals $\hat{T}_1,\cdots,\hat{T}_p$

$$\hat{\theta} = h(\hat{T}_1,...,\hat{T}_p) = a_0 + \sum_{j=1}^{p} a_j \hat{T}_j$$

The variance of $\hat{\theta}$ may be written as

$$V(\hat{\theta}) = V\left(\sum_{j=1}^{p} a_j \hat{T}_j\right) = \sum_{j=1}^{p}\sum_{l=1}^{p} a_j a_l \; cov(\hat{T}_j, \hat{T}_l), \tag{4.3.1}$$

where the covariance between $\hat{T}_j$ and $\hat{T}_l$, $cov(\hat{T}_j, \hat{T}_l)$, is equal to variance of $\hat{T}_j$ for $j = l$. The variance (4.3.1) can be easily estimated by using estimated covariance terms as illustrated in Särndal et al. (1992, page 172).

However, in the case of h being a nonlinear function of the p totals, it is often impossible to obtain an exact expression for the sampling variance of the estimator $\hat{\theta} = h(\hat{T}_1,...,\hat{T}_p)$. Then, the Taylor linearization method may be used to obtain an approximate expression for the variance of $\hat{\theta}$ and also an approximate estimator of this variance. This method approximates the nonlinear estimator $\hat{\theta}$ by a pseudo-estimator $\hat{\theta}_0$, which is a linear function of $\hat{T}_1,...,\hat{T}_p$ and thus easy to manipulate. The technique for finding $\hat{\theta}_0$ consists of the first-order Taylor approximation of the function h, expanding around the point $(T_1,\cdots,T_p)$, defined as the expectation of $(\hat{T}_1,\cdots,\hat{T}_p)$, and neglecting the remainder term. That is

$$\hat{\theta} \approx \hat{\theta}_0 = \theta + \sum_{j=1}^{p} a_j (\hat{T}_j - T_j), \tag{4.3.2}$$

where $a_j = \left.\frac{\partial h}{\partial \hat{T}_j}\right|_{(\hat{T}_1,...,\hat{T}_p)=(T_1,...,T_p)}$.

When $\hat{T}_1,...,\hat{T}_p$ with high probability take values near $T_1,\cdots,T_p$, the estimator $\hat{\theta}$ performs approximately as the linear random variable $\hat{\theta}_0$. The numeric accuracy of the approximation (4.3.2) will vary from one outcome s to another. Finally, the variance of $\hat{\theta}$ can be approximated by the corresponding derived quantities for the linear statistics $\hat{\theta}_0$

$$V(\hat{\theta}) \approx V(\hat{\theta}_0) = V\left(\sum_{j=1}^{p} a_j \hat{T}_j\right). \tag{4.3.3}$$

Now consider the application of the Taylor linearization method to the weighted estimators $\hat{T}_y$ defined in section 4.2. This weighted estimator may be expressed as a function of estimated totals. For example, if $\hat{\theta}$ denotes the generalised regression estimator as defined in (4.2.7), then $p = 4$, $\hat{T}_1 = \hat{T}_{yd}$, $\hat{T}_2 = \hat{T}_{xd}$, $\hat{T}_3 = \sum_r d_i \boldsymbol{x}_i \boldsymbol{x}_i'$ and $\hat{T}_4 = \sum_r d_i \boldsymbol{x}_i y_i$. A nonresponse mechanism is assumed such that each unit in the population responds, if sampled, with probability q_i, where this probability is not dependent on the choice of the sample and different units respond independently. Therefore, the response mechanism is viewed as a second phase of sampling and the variance is defined with respect to the joint distribution induced by both sampling and nonresponse.

It is important to note that in general the class of weighted estimators presented in section 4.2 (and in particular the classical and the "maximum likelihood" raking) involves iterative modifications of the initial weights d_i to calibrated weights w_i with the aim of satisfying the calibration equations (4.2.2). Following Binder and Théberge (1988) and Deville et al. (1993), this section seeks to estimate the asymptotic variance of the 'converged' estimator, i.e. the estimator $\hat{T}_y$, where the w_i are the 'converged' weights that solve the calibration equations. Some research exits on estimating the variance of $\hat{T}_y$ after a finite number of iterations (Deville et al., 1993).

The nonlinear nature of the weighted estimator $\hat{T}_y$ in (4.2.1) arises through the weights w_i and their dependence on $\hat{\lambda}$ via expression (4.2.5). It is assumed that in large samples, $\hat{\lambda}$ converges to a value λ. Deville and Särndal (1992) assumed that $\lambda = \mathbf{0}$, but this property is based upon the assumption that the estimator of T_x, obtained by applying the initial weights d_i is consistent. This assumption will often be false in the case of nonresponse and thus it is not made in this chapter.

A linearization variance estimator is obtained by approximating $\mathrm{var}(\sum_r w_i y_i)$ by $\mathrm{var}(\sum_r d_i z_i)$ for a 'linearized variable' z_i (Deville 1999). First, an expression for $\hat{\lambda}$ is obtained. A Taylor expansion of the calibrated weight $w_i = d_i F(\boldsymbol{x}_i'\hat{\lambda})$ about λ results in

$$w_i \approx \boldsymbol{d}_i \left[F_i + f_i \boldsymbol{x}_i'(\hat{\lambda} - \lambda) \right], \tag{4.3.4}$$

where $\approx$ denotes 'is asymptotically equivalent to', $F_i = F(\boldsymbol{x}_i'\lambda)$, $f_i = f(\boldsymbol{x}_i'\lambda)$, and $f\ u\ = dF\ u\ /du$ is assumed to exist. Substituting in the calibration equations we obtain:

$$\sum_{i\in r} w_i \boldsymbol{x}_i \approx \sum_{i\in r} \boldsymbol{d}_i \left[F_i + f_i \boldsymbol{x}_i'(\hat{\lambda} - \lambda)\right]\boldsymbol{x}_i \approx \boldsymbol{T}_x,$$

and hence

$$\hat{\lambda} - \lambda \approx \left[\sum_{i\in r} \boldsymbol{d}_i f_i \boldsymbol{x}_i \boldsymbol{x}_i'\right]^{-1} \left[\boldsymbol{T}_x - \sum_{i\in r} \boldsymbol{d}_i F_i \boldsymbol{x}_i\right]. \tag{4.3.5}$$

The first matrix in the expression (4.3.5) is assumed non-singular. It may be necessary to drop redundant variables from $\boldsymbol{x}_i$ to achieve this. For example, in the three-way case within the context of the Labour Force Survey presented in section 4.2, each of the sums of the indicator variables $\delta_{a..i}$, $\delta_{.b.i}$ and $\delta_{..ci}$ across a , b and c , respectively, equals 1 and it is natural to drop two of these indicators to avoid singularity. The non-singular condition might also require (as in Deville and Särndal, 1992) modifying the estimator for samples with small probability.

Substituting in the calibrated estimator results in:

$$\hat{T}_y \approx \sum_{i\in r} \boldsymbol{d}_i \left[F_i + f_i \boldsymbol{x}_i'(\hat{\lambda} - \lambda)\right] y_i \approx \sum_{i\in r} d_i F_i y_i + B\left[\boldsymbol{T}_x - \sum_{i\in r} d_i F_i \boldsymbol{x}_i\right], \tag{4.3.6}$$

where

$$B = \left[\sum_{i\in r} d_i f_i y_i \boldsymbol{x}_i'\right]\left[\sum_{i\in r} \boldsymbol{d}_i f_i \boldsymbol{x}_i \boldsymbol{x}_i'\right]^{-1}. \tag{4.3.7}$$

Note that $F_i = f_i = 1$ under the assumptions of Deville and Särndal (1992) (since in this case $\lambda = 0$ and it follows from the assumptions about $G(.)$ that $F(0) = f(0) = 1$). Hence, under these assumptions, expression (4.3.6) corresponds to Result 5 of Deville and Särndal (1992), i.e. the generalized raking estimator is asymptotically equivalent to the GREG estimator. Therefore, the asymptotic variance of $\hat{T}_y$ is the same as that of $\sum_{i\in r} d_i z_i$, where z_i is the linearized variable

$$z_i = F_i\ y_i - \beta \boldsymbol{x}_i\ , \tag{4.3.8}$$

assuming that B converges to a finite limit vector β .

For the purpose of linearization variance estimation and following the derivation above, $\hat{T}_y$ may be treated as the linear estimator $\sum_{i\in r} d_i \hat{z}_i$, where

$$\hat{z}_i = \hat{F}_i(y_i - \hat{B}\boldsymbol{x}_i), \tag{4.3.9}$$

and $\hat{z}_i$ is obtained by replacing the unknown parameters in (4.3.8) by the later discussed estimators $\hat{F}_i$ and $\hat{B}$. Then, having determined $\hat{z}_i$, the linearization variance estimator for $\hat{T}_y$ for a given sampling design is obtained by using a standard variance estimator for that design for a linear estimator, applied to $\sum_{i\in r} d_i \hat{z}_i$.

For example, in the case of a stratified multistage sampling design, assuming "with replacement" sampling of primary sampling units (PSUs) within strata, a standard estimator of the variance of $\hat{T}_y$ (e.g. Stukel et al., 1996) is given by:

$$\hat{V}(\hat{T}_y) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (z_{hj} - \overline{z}_h)^2, \tag{4.3.10}$$

where $z_{hj} = \sum_k d_{hij}\hat{z}_{hjk}$, $\overline{z}_h = \sum_j z_{hj}/n_h$, and $\hat{z}_{hjk}$ is the value of the variable defined in (4.3.9) for the kth individual within the jth selected PSU in stratum h. This estimator remains appropriate in the presence of nonresponse if individual response in each PSU is independent of response in all other PSUs and if at least one individual is observed in each selected PSU (Fuller et al., 1994, p.78).

In order to obtain $\hat{z}_i$, a number of choices for $\hat{F}_i$ and $\hat{B}$ are considered in the literature. Regarding $\hat{F}_i$, a natural choice implied by the above argument would be to select $\hat{F}_i$ such as $\hat{F}_i = F(\boldsymbol{x}_i'\hat{\lambda})$. This choice results in the linear estimator written as:

$$\sum_{i\in r} d_i \hat{z}_i = \sum_{i\in r} d_i F(\boldsymbol{x}_i'\hat{\lambda})(y_i - \hat{B}\boldsymbol{x}_i) = \sum_{i\in r} w_i(y_i - \hat{B}x_i). \tag{4.3.11}$$

This chapter will refer to (4.3.11) as the w_i*-weighted residuals* estimator. Another simpler choice for $\hat{F}_i$ would be $\hat{F}_i = 1$, which leads to the d_i - *weighted residuals* estimator:

$$\sum_{i\in r} d_i \hat{z}_i = \sum_{i\in r} d_i(y_i - \hat{B}\boldsymbol{x}_i). \tag{4.3.12}$$

Deville and Särndal (1992) noted that, in their classical theory with $\lambda = 0$, both choices are asymptotically equivalent. However, they expressed a preference for the choice

$\hat{F}_i = F(\boldsymbol{x}_i'\hat{\lambda})$. This preference is also highlighted by Fuller (2002, p.15), in particular, within the nonresponse setting and with $\lambda = 0$ not necessarily holding as in this study.

Regarding $\hat{B}$, it follows from the argument on the choices of $\hat{F}_i$ that f_i in (4.3.7) should be replaced by $\hat{f}_i = f(\boldsymbol{x}_i'\hat{\lambda})$, giving:

(i) $\hat{B} = \left[\sum_r d_i \hat{f}_i y_i x_i'\right]\left[\sum_r d_i \hat{f}_i x_i x_i'\right]^{-1}$, as also proposed by Demnati and Rao (2004).

Other choices are

(ii) $\hat{B} = \hat{B}_r$, as in (4.2.8), as proposed by Deville et al. (1993).

(iii) $\hat{B} = \left[\sum_r w_i y_i x_i'\right]\left[\sum_r w_i x_i x_i'\right]^{-1}$, as proposed by Deville and Särndal (1992, equation 3.4), which might be more practical to compute than $\hat{B}_r$ for users of survey data files which include the w_i weights but not the d_i weights.

The extent to which these choices differ depends on the choice of the $G(.)$ function. For the linear case $f(u) = dF\ u\ / du = d(1+u)/du = 1$ so that the estimators in (i) and (ii) are identical. In the case of classical raking adjustment, $f(u) = d\left[\exp(u)\right]/du = \exp(u) = F(u)$ so that $\hat{f}_i = \hat{F}_i$ and $d_i\hat{f}_i = d_i F(\boldsymbol{x}_i'\hat{\lambda}) = w_i$ and the estimators (i) and (iii) are identical. For the 'maximum likelihood' raking estimator we have $F(u) = (1-u)^{-1}$ and $f(u) = (1-u)^{-2}$ so that $d_i\hat{f}_i = w_i^2 / d_i$ and the three variance estimators are all distinct.

4.4 Replication variance estimation

Another class of methods used for computing sampling variance estimators for nonlinear survey statistics is *subsample replication* (Wolter, 2007). These methods derive estimates of the parameter of interest from each of several subsamples of the original sample and then estimate the variance of the original sample estimator from the variability between the subsample estimates. Following the notation in section 4.2, a replication estimator of the variance of $\hat{T}_y$ may be obtained by: (1) constructing a set of

A replicate weights $w_i^{(a)}$ for $a = 1, \ldots, A$, using different replicate sampling technique for the different replication methods; (2) computing for each set of replicate weights an estimator $\hat{T}_y^{(a)}$ of T_y in the same way that $\hat{T}_y$ is computed using the weights w_i; (3) using the A replicate estimates and the original sample estimate, compute the estimator of the variance of $\hat{T}_y$ using the following equation:

$$\hat{V}(\hat{T}_y) = \sum_{a=1}^{A} c_a (\hat{T}_y^{(a)} - \hat{T}_y)^2 , \tag{4.4.1}$$

where c_a is a constant which depends on the replication method.

The construction of the replicate weights $w_i^{(a)}$ involves first taking the initial weights d_i and constructing from these a set of initial replication weights $d_i^{(a)}$, $a = 1, \ldots, A$, according to the replication method and the sampling scheme. Then, calibration adjustments are applied to each of these A sets of initial weights separately.

One frequently used replication technique to calculate variance estimators is the *jackknife* method. In a stratified multistage cluster sampling design, this method is applied separately in each stratum at the first stage of sampling, with one primary sampling unit (PSU) deleted at a time. The number of replicates in this case is $A = \sum n_h$, where n_h is the number of PSUs in stratum $h = 1, \ldots, H$ and H is the number of strata in the population. To apply the jackknife, let replicate a correspond to deleting PSU j in stratum h, calculate the replicate initial weights

$$d_i^{(a)} = \begin{cases} d_i & \text{if observation unit } i \text{ is not in stratum } h. \\ d_i / c_a & \text{if observation unit } i \text{ is in stratum } h \text{ but not in PSU } j. \\ 0 & \text{if observation unit } i \text{ is in PSU } j \text{ of stratum } h. \end{cases} \tag{4.4.2}$$

where $c_a = n_h - 1 \,/\, n_h$, for $a = 1, \ldots, A$. Then use the weights $d_i^{(a)}$ to compute $w_i^{(a)}$ using generalized raking estimation as in section 4.2. Finally, calculate $\hat{T}_y^{(a)}$ and the variance estimator of $\hat{T}_y$ using (4.4.1).

The jackknife method described above requires that each cluster within each stratum is deleted in turn. This could require many recalculations for large surveys and thus be prohibitive. An alternative is to group the n_h clusters in the hth stratum into $g_h \geq 2$ groups ($g_h < n_h$) and to proceed as if these were the actual clusters (see, for

example, Valliant, 2004). Thus each group is deleted in turn and the number of recalculations is reduced to $g = \sum_h g_h$. The only change in (4.4.1) is in c_a as follows:

$$c_a = \frac{(g_h - 1)}{g_h}. \tag{4.4.3}$$

Another subsample replication method for variance estimation is the *bootstrap* method. The idea of the bootstrap method is to use the variance in repeated bootstrap sampling to estimate the variance of the point estimator. A bootstrap sample is a simple random sample with replacement of size n, for example, selected from the original sample. The bootstrap estimator of the variance of $\hat{T}_y$ may be obtained by (4.4.1) where $c_a = 1/A$. This method is computationally more intensive than the Jackknife method and is not further investigated in this chapter. For more details about alternative bootstrap method and applications see Wolter (2007).

Section 4.6 reports a simulation study of the properties of group jackknife variance estimators of the generalised regression estimator introduced in section 4.2.

4.5 Simulation studies

In order to compare the performance of the weighted estimators presented in section 4.2 and their corresponding variance estimators, discussed in section 4.3 and 4.4, two simulation studies are undertaken by constructing artificial populations using data from the Great Britain Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). In each case, $R = 1{,}000$ samples are generated from these populations by first sampling, in a way designed to mimic as far as possible the real sampling scheme after some simplification, and then removing nonresponding cases according to two nonresponse models. This study shall refer to the first model as the multiplicative nonresponse model and to the second as the additive nonresponse model.

For every one of the R samples, point estimates of each of the parameters are calculated using generalized raking estimation and variance estimates are computed using linearization and replication methods. The properties of the estimators are then summarised.

4.5.1 Study based on the British Labour Force Survey

The British LFS is a quarterly survey of persons living in private households in Britain. Its purpose is to provide information on the British labour market which can then be used to develop, manage, evaluate and report on labour market policies. It is carried out by the Social Survey Division of the Office for National Statistics (ONS).

Artificial population and sampling design

In the first simulation study, data from the March-May 1998 quarter of the British LFS is treated as an artificial population. The LFS is a very large survey which results in approximately 58,000 addresses in the artificial population. From this population repeated samples were drawn in a way intended to mimic as far as possible the design used for the LFS. Details about the design of the survey can be found in ONS (1998, Section 3). Each sample consists of 1211 households (cluster of individuals) selected by stratified simple random sampling with proportional allocation across 19 strata, defined by region of residence. These regions are used to mimic the effect of the 110 Interviewer Areas (IAs) which defined strata in the LFS. In the LFS *all* individuals in a sampled household are interviewed if possible. In this simulation study, all the respondents in a sample household are retained, except those aged under 16, who are not relevant for the estimates of interest.

Unit nonresponse

Nonresponse probabilities are assigned to each household in the generated artificial population. It is assumed in the simulation study that all individuals within a household respond. Two different nonresponse models are considered to determine whether sampled households respond, a multiplicative and an additive model. Information to assign nonresponse probabilities to each selected household of the artificial population is obtained from a study of Foster (1998), in line with findings in Chapter 2, and takes into account characteristics of households, such as area of residence, age and gender of household reference person (HRP). It is assumed that household nonresponse depends on these auxiliary variables but not on the survey variables of interest, which is similar to assume a missing at random mechanism for nonresponse (Little and Rubin, 1987).

Multiplicative Nonresponse Model:

q_i^{-1} = 1.15 × 1.17 (if London) × 1.13 (if HRP aged under 35) × 1.10 (if HRP female)

Additive Nonresponse Model:

q_i^{-1} = 1.15 + 0.20 (if London) + 0.15 (if HRP aged under 35) + 0.10 (if HRP female)

where q_i is the response probability for each household i in the population, if sampled. The response probability is not dependent on the choice of the sample and different households respond independently. Kott (2006) and Chang and Kott (2008) consider estimating response probabilities using general models of the form $q_i^{-1} = F(x_i'\alpha)$ (also see Skinner and D'Arrigo, 2011, Sect. 3). The first model assumes multiplicative nonresponse, which might be expected to lead to least bias for the raking ratio method (see, for example, D'Arrigo and Skinner, 2010, Section 3), and the second model assumes additive nonresponse, which might be expected to lead to least bias for the GREG estimator (see, for example, Fuller, 2002, Section 8). Therefore, these models are designed so that the raking and GREG estimators respectively perform well.

Weighting and Calibration

Weights are constructed for responding individuals within selected households, with calibration totals consisting of population counts in the categories of three categorical auxiliary variables: area of residence, age and gender, and with Horvitz-Thompson initial weights d_i , as in section 4.2. The choice of auxiliary variables was designed to mimic those used in the LFS. However, because of small numbers of individuals within strata due to our artificial population and samples being much smaller than those for the original survey, we simplified the LFS calibration variables to the following three categorical factors:

- area of residence (see Appendix A3) with 23 categories;
- a cross-classification of gender by age groups (with 10 age groups consisting of single years for those between 16 and 24 and a separate age group for those 25 or older) with 20 categories;
- a cross-classification of region (Northern England; London and South East; The Midlands and East Anglia; Scotland) by gender by age groups (in 15-year age groups: 16-29, 30-44, 45-59, 60-75 and 75 or older), with 40 categories.

Survey statistics

The parameters of interest defined for the artificial population are: the total number of persons unemployed (TNU), the total number of persons employed (TNE) and the total number of persons in the inactive workforce (TNI). These parameters are computed using the artificial population of the LFS by

$$T_y = \sum_{(jih)\in U} y_{jih}$$

where y_{jih} denotes the vector of indicator survey variables: $y_{jih} = y_{1jih}, y_{2jih}, y_{3jih}$, where $y_{1jih} = 1$ if individual j in household i within stratum h is unemployed and 0 otherwise, $y_{2jih} = 1$ if individual j in household i within stratum h is employed and 0 otherwise, and $y_{3jih} = 1$ if individual j in household i within stratum h is inactive and 0 otherwise.

For each of the R simulated samples, point estimates of the TNU, TNE and TNI are computed using classical raking estimation, "maximum likelihood" raking estimation, and generalized regression estimation and, associated variance estimates are calculated using the alternative linearization methods described in section 4.3 and the jackknife replication technique illustrated in section 4.4. The properties of these estimators, under alternative assumptions about nonresponse, are investigated following usual practice in simulation studies. For example:

(1) The bias of the point estimator $\hat{T}_y$ with respect to the population parameter T_y is estimated by:

$$Bias(\hat{T}_y) = \frac{1}{R}\sum_{r=1}^{R}(\hat{T}_{y_r} - T_y),$$

where $\hat{T}_{y_r}$ is the value of $\hat{T}_y$ for sample r.

(2) The percent relative bias of the point estimator $\hat{T}_y$ with respect to the population parameter is estimated by:

$$\frac{Bias(\hat{T}_y)}{T_y} * 100 .$$

(3) The simulation variance of the point estimator $\hat{T}_y$ taken over the R samples is estimated by:

$$V_s = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{T}_{y_r} - E(\hat{T}_y)\right]^2,$$

where $E(\hat{T}_y) = \frac{1}{R}\sum_{r=1}^{R}\hat{T}_{y_r}$.

(4) The simulation variance of the bias estimator (1) from:

$$V\left[Bias(\hat{T}_y)\right] = \frac{V_s}{R}.$$

(5) The root mean square error of the point estimator $\hat{T}_y$ is estimated by:

$$RMSE(\hat{T}_y) = \sqrt{V_s + \left[Bias\ \hat{T}_y\right]^2}.$$

(6) The expectation of the variance estimator of $\hat{T}_y$ taken over the R samples from:

$$E\left[\hat{V}(\hat{T}_y)\right] = \frac{1}{R}\sum_{r=1}^{R}\hat{V}_r(\hat{T}_y),$$

where $\hat{V}_r(\hat{T}_y)$ is the value of the variance estimate for sample r.

(7) The bias of the variance estimator of $\hat{T}_y$ with respect to the simulation variance (3) is estimated by:

$$Bias\left[\hat{V}(\hat{T}_y)\right] = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{V}_r(\hat{T}_y) - V_s\right].$$

(8) The percent relative bias of the variance estimator of $\hat{T}_y$ with respect to the simulation variance (3) is estimated by:

$$\frac{Bias\left[\hat{V}(\hat{T}_y)\right]}{V_s} * 100.$$

(9) The variance of the bias of the variance estimator of $\hat{T}_y$ from:

$$V\ \ Bias\left[\hat{V}(\hat{T}_y)\right]\ \ = \frac{V\left[\hat{V}(\hat{T}_y)\right]}{R},$$

where $V\left[\hat{V}(\hat{T}_y)\right] = \frac{1}{R}\sum_{r=1}^{R}\left\{\hat{V}_r(\hat{T}_y) - E\left[\hat{V}(\hat{T}_y)\right]\right\}^2$.

(10) The root mean square error of the variance estimator of $\hat{T}_y$ from:

$$RMSE\left[\hat{V}(\hat{T}_y)\right] = \sqrt{V\left[\hat{V}(\hat{T}_y)\right] + \ Bias\left[\hat{V}(\hat{T}_y)\right]^2}.$$

(11) A confidence interval for the population parameter T_y for sample r at the approximate 95% level, defined as:

$$\hat{T}_{y_r} \pm 1.96\sqrt{\hat{V}_r(\hat{T}_y)}.$$

In order to check if this confidence interval is valid, that is, if the desired 95% normal-theory confidence level is attained, an empirical validation is carried out by simulation. First, for each sample $r = 1,\ldots,R$, the estimator $\hat{T}_{y_r}$, the variance estimator $\hat{V}_r(\hat{T}_y)$, and the confidence interval defined above are computed. Then, for each of the R confidence intervals computed, observe whether the known parameter T_y is included in the interval or not. If K of the R intervals are found to contain T_y, the empirical coverage of the confidence interval is defined as the proportion H/R. This proportion should lie near the desired 95% confidence level.

Some statistics related to the calibration weights, such as the number of negative weights and number of weights more than 10 times the corresponding design weights, resulting from using each of the function $G(.)$ under study are also computed.

4.5.2 Study based on the German Sample Survey of Income and Expenditure

The Sample Survey of Income and Expenditure (SIE) is a nationwide household survey conducted every 5 years by the Federal Statistical Office. The main purpose of the survey is to provide information about the economic and social situation of households, in particular regarding the distribution of income and expenditure (Quatember et al., 2002).

Population and Sampling Design

The second simulation study is based on the 1998 SIE. It uses data from an artificial population of 64,326 households, created to represent 20% of all households from the Bremen region, excluding those with a monthly household net income of DM 35,000 or above (DM denotes the currency German marks). The SIE employs a quota sampling design which is not attempted to mimic in this simulation study. Instead, simple random sampling allowing for nonresponse is employed in this simulation. Repeated simple random samples of 1340 households are drawn from the synthetic population, representing a sampling fraction of about $1/48$.

Unit nonresponse

Even though the SIE quota sampling design does not allow for nonresponse, two different nonresponse models are considered in this study for research purposes, a multiplicative and an additive. Information to assign nonresponse probabilities to each selected household from the artificial population is obtained from results of studies of similar surveys in Great Britain: the Family Expenditure Survey and the National Food Survey (Foster, 1998). This information takes into account characteristics of households, such as socio-economic status and type of household. For each selected sample, the subset of responding households is determined by the following nonresponse models:

Multiplicative Model:

q_i^{-1} = 1.44 × 1.09 (if HRP self-employed) × 1.03 (if HRP unemployed) × 0.97 (if HRP employed) × 1.16 (if no children in the household)

Additive Model:

q_i^{-1} = 1.44 + 0.13 (if HRP self-employed) + 0.04 (if HRP unemployed) – 0.04 (if HRP employed) + 0.23 (if no children in the household)

Weighting and Calibration

As for the LFS study, each sampled household is assigned a weight. In the actual SIE the weights are constructed using essentially the maximum likelihood raking method by adjusting the sample data simultaneously to the marginal distributions of several characteristics, such as household type, social economic status of the household reference person, household net income class and region (Bundesland). This study tries

to mimic this adjustment as far as possible. However, as for the LFS study, because of the reduced scale of the created artificial population and the consequent smaller numbers of households within strata, the SIE calibration variables are simplified to three categorical factors as follows:

- household type with 7 categories (mother/father alone + 1 child; mother/father alone + 2 or more children; couple with 1 child - spouse employed; couple with 1 child - spouse unemployed; couple with 2 or more children - spouse employed; couple with 2 or more children - spouse unemployed; other);
- social status of the household reference person with 5 categories (self-employed; civil servant or military; employee; worker; unemployed, pensioner, student or other);
- household net income per quarter with 3 categories (0-5,000 DM; 5-7,000 DM; 7-35,000 DM).

Survey statistics

The parameters of interest are the total household net income per quarter (INC) and the total household expenditure per quarter (EXP). These parameters are computed from the finite artificial population by

$$T_{y_{INC}} = \sum_{i \in U} y_i \qquad \text{and} \qquad T_{z_{EXP}} = \sum_{i \in U} z_i \,,$$

where y_i and z_i denote the value of the continuous survey variable INC and EXP for household i, respectively.

As in the LFS study, for each of the R samples, point estimates of the above parameters and associated variance estimates are calculated using the different methods presented in this chapter. The properties of the estimators are then summarised in the following section.

4.6 Results

4.6.1 Properties of point estimators

Table 4.6.1 presents the properties of the point estimators of total number of persons unemployed in the LFS study for different calibration methods and alternative

assumptions about nonresponse. It also shows the number of negative weights and very large weights for the different settings and across all sample units and all 1000 samples. Numbers are rounded to the nearest one decimal. An overall observation from this table is that the standard error remains virtually constant across alternative raking methods for a given nonresponse model. As expected, nonresponse leads to an increase in the standard error across all estimators (since the sample size is reduced). Regarding bias, the table shows evidence of nonresponse bias relative to the simulation standard error of bias ($z = bias/se(bias) > 1.96$), which is of a similar order for each of the raking methods. It is not found that this bias is least when the estimator matches the nonresponse model (i.e. the GREG estimator for additive response and the raking estimator for multiplicative response) as it might have expected. Perhaps this is because the covariates used in the nonresponse models (e.g. the aged 35+ variable) are not all included in the calibrating variables. Nevertheless, the nonresponse bias is small (relative bias of about 1% across weighting methods) in the sense that the root mean square error is very similar to the standard error in each case. Under nonresponse, the GREG calibration method generates some negative weights whereas this is avoided by the two raking methods, as expected. A greater number of very large weights are observed, however, for the 'maximum likelihood' raking estimator.

Table 4.6.1: Simulation properties of point estimators of total unemployed using data from LFS (R=1000)

Nonresponse Model/ Point Estimator	Bias (simulation standard error)	Percent Relative Bias	Standard Error	Root Mean Square Error	Number of Negative Weights[1]	Number of Very Large Weights[1,2]
Complete Response:						
GREG	7.6 (14.3)	0.2	452.8	452.8	0	0
Classical Raking	8.3 (14.3)	0.2	452.8	452.9	0	0
'ML' Raking	9.0 (14.3)	0.2	453.3	453.4	0	1
Multiplicative nonresponse:						
GREG	-45.6 (15.8)	-1.2	498.3	500.3	4	1
Classical Raking	-42.1 (15.8)	-1.1	498.8	500.6	0	2
'ML' Raking	-39.7 (15.8)	-1.0	499.4	501.0	0	7
Additive nonresponse:						
GREG	-37.3 (15.7)	-0.9	497.4	498.8	5	1
Classical Raking	-34.7 (15.7)	-0.9	497.5	498.7	0	3
'ML' Raking	-32.4 (15.8)	-0.8	498.1	499.1	0	7

[1]the number of such weights across all sample units and all 1000 samples
[2]the number of weights more than 10 times the corresponding design weight

Corresponding results for the SIE data are presented in Table 4.6.2. The pattern of results is broadly similar, although there is now no evidence of significant nonresponse bias (i.e. the observed bias could be explained by simulation variation). The standard errors and root mean square errors also remain virtually constant across weighting methods for a given nonresponse model. There are no negative weights or very large weights observed in Table 4.6.2.

Table 4.6.2: Simulation properties of point estimators of total income using data from SIE (R=1000)

Nonresponse Model/ Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negative Weights[1]	Number of Very Large Weights[1,2]
Complete Response:					
GREG	-172.2 (331.3)	10,477.3	10,478.7	0	0
Classical Raking	-170.6 (331.5)	10,484.1	10,485.8	0	0
'ML' Raking	-169.8 (331.8)	10,491.5	10,492.9	0	0
Multiplicative nonresponse:					
GREG	-495.7 (429.7)	13,586.8	13,595.8	0	0
Classical Raking	-493.8 (429.6)	13,584.6	13,593.5	0	0
'ML' Raking	-463.5 (429.5)	13,582.8	13,590.7	0	0
Additive nonresponse:					
GREG	-473.2 (430.5)	13,614.8	13,623.0	0	0
Classical Raking	-469.4 (430.5)	13,612.9	13,621.0	0	0
'ML' Raking	-439.5 (430.5)	13,613.5	13,620.6	0	0

[1]the number of such weights across all sample units and all 1000 samples
[2]the number of weights more than 10 times the corresponding design weight

4.6.2 Properties of variance estimators

The properties of the different linearization estimators of the variances of the point estimators of the total unemployed from the LFS are shown in the Table 4.6.3 (the 'standard error estimate' in the table refers to the square root of the variance estimate). A number of observations can be made from this table as follows:

- Using calibrated weights w_i to weight the residuals rather than using initial weights d_i, reduces the bias and root mean squared error of the standard error estimator. This is observed across all alternative raking methods and nonresponse assumptions. The bias arising from the use of d_i weighted residuals in the case of nonresponse is particularly important (as noted by Fuller, 2002) but there are also non-negligible reductions of bias even in the complete response case.

- The choice of weight used in the estimated regression coefficients $\hat{B}$ for the calculation of residuals seems to have little impact. Some slight evidence in favour of using initial weights d_i to compute $\hat{B}$ when simultaneously weighting the residuals by calibrated weights w_i might be observed.

- For a given nonresponse setting and choice of weighting the residuals, there is little difference in the results for the different choices of point estimator.

The results in Table 4.6.3 are extended in Table 4.6.4 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider two additional parameters: total numbers employed and inactive. From table 4.6.4, it can be again observed that the relative bias arising from using d_i weighted residuals can be substantial in the presence of nonresponse, over 20% in several cases, and that this is reduced using the w_i weighted residuals. Again, little change is observed in the percent relative bias of the standard error estimators when different choices of weights are used in the calculation of $\hat{B}$ for the residuals. It is important to note that confidence interval coverages and relative biases reported in Table 4.6.3 and 4.6.4 respectively are not expected to be affected by the small nonresponse bias in the estimates of the totals.

Corresponding results for the SIE data when estimating total income are shown in Table 4.6.5. Again, the pattern of results is broadly similar to that for the LFS data in Table 4.6.3. For the complete response case, the use of w_i-weighted residuals rather than d_i-weighted residuals leads to modest improvement in bias and RMSE of the standard error estimators. However, for the nonresponse cases the improvements are considerable. Little change in the standard error estimators is observed when modifying the choice of weight used to compute the estimated regression coefficients, observing again slightly smaller biases when using initial weights to compute $\hat{B}$ and calibrated weights to weight the residuals. However, the results do not suggest that one approach leads to consistently lower absolute bias. The results in Table 4.6.5 are extended in Table 4.6.6 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider one additional parameter: total expenditure per quarter. It is again observed that the relative bias arising from using d_i-weighted residuals can be substantial in the presence of nonresponse, over 35% in all cases, and that this is reduced using the w_i-weighted residuals, for which the relative bias never exceeds about 3%.

Table 4.6.3: Properties of linearization variance estimators when estimating total unemployed from the LFS (R = 1000)

Weighting Method	w - or d - weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	d	d	433.9	-18.8 (0.9)	33.4	93.5
	d	w	434.3	-18.5 (0.9)	33.3	93.5
	w	d	442.8	**-10.0 (1.0)**	**31.9**	93.8
	w	w	441.9	**-10.8 (1.0)**	**32.0**	93.7
Classical Raking	d	d	433.9	-18.8 (0.9)	33.4	93.5
	d	w	434.2	-18.5 (0.9)	33.3	93.5
	w	d	443.0	**-9.8 (1.0)**	**32.0**	93.8
	w	w	442.0	**-10.7 (1.0)**	**32.0**	93.8
'ML' Raking	d	d	433.9	-19.4 (0.9)	33.7	93.5
	d	w	434.3	-19.1 (0.9)	33.6	93.5
	d	df	435.4	-17.9 (0.9)	33.0	93.5
	w	d	443.7	**-9.6 (1.0)**	**32.5**	93.7
	w	w	442.3	**-11.1 (1.0)**	**32.4**	93.7
	w	df	441.6	**-11.8 (1.0)**	**32.3**	93.7
Multiplicative nonresponse:						
GREG	d	d	385.7	-112.6 (0.9)	116.0	85.8
	d	w	386.1	-112.1 (0.9)	115.5	85.8
	w	d	489.5	**-8.8 (1.2)**	**39.2**	94.2
	w	w	487.8	**-10.4 (1.2)**	**39.2**	94.2
Classical Raking	d	d	385.7	-113.1 (0.9)	116.5	85.7
	d	w	386.1	-112.7 (0.9)	116.1	85.7
	w	d	490.3	**-8.5 (1.2)**	**39.6**	94.3
	w	w	488.4	**-10.4 (1.2)**	**39.5**	94.1
'ML' Raking	d	d	385.7	-113.7 (0.9)	117.1	85.4
	d	w	386.2	-113.2 (0.9)	116.6	85.6
	d	df	387.8	-111.6 (0.9)	115.0	85.8
	w	d	491.9	**-7.5 (1.3)**	**40.4**	94.2
	w	w	488.9	**-10.5 (1.2)**	**39.9**	94.0
	w	df	487.5	**-11.9 (1.2)**	**39.8**	94.0
Additive nonresponse:						
GREG	d	d	386.5	-110.9 (0.9)	114.4	86.0
	d	w	387.0	-110.5 (0.9)	113.9	86.0
	w	d	489.3	**-8.2 (1.2)**	**39.0**	94.6
	w	w	487.6	**-9.8 (1.2)**	**39.0**	94.6
Classical Raking	d	d	386.5	-111.0 (0.9)	114.4	85.8
	d	w	387.0	-110.6 (0.9)	114.0	85.8
	w	d	490.1	**-7.4 (1.2)**	**39.2**	94.7
	w	w	488.1	**-9.4 (1.2)**	**39.1**	94.6
'ML' Raking	d	d	386.5	-111.6 (0.9)	115.0	85.6
	d	w	387.0	-111.1 (0.9)	114.6	85.6
	d	df	388.6	-109.5 (0.9)	113.0	85.9
	w	d	491.6	**-6.5 (1.3)**	**40.0**	94.7
	w	w	488.6	**-9.5 (1.2)**	**39.5**	94.6
	w	df	487.3	**-10.8 (1.2)**	**39.4**	94.6

Figures in bold indicate the best approach under each scenario

[1] see text following equation (4.3.11), where choices df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively

[2] percentage of 95% normal-theory confidence intervals containing true value

Table 4.6.4: Percent relative bias of linearization standard error estimators of unemployed, employed and inactive totals from LFS (R = 1000)

Weighting Method	w - or d - weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Percent Relative Bias of Standard Error Estimator		
			Unemployed	Employed	Inactive
Complete Response:					
GREG	d	d	-4.2	-3.4	0.5
	d	w	-4.1	-3.3	0.6
	w	d	-2.2	-2.2	1.9
	w	w	-2.4	-2.3	1.7
Classical Raking	d	d	-4.2	-3.3	0.7
	d	w	-4.1	-3.2	0.8
	w	d	-2.2	-2.1	2.1
	w	w	-2.4	-2.2	1.9
'ML' Raking	d	d	-4.3	-3.3	0.7
	d	w	-4.2	-3.3	0.8
	d	df	-4.0	-3.1	1.1
	w	d	-2.1	-2.0	2.3
	w	w	-2.4	-2.2	1.9
	w	df	-2.6	-2.3	1.8
Multiplicative nonresponse:					
GREG	d	d	-22.6	-22.3	-18.2
	d	w	-22.5	-22.2	-18.1
	w	d	**-1.8**	**-3.3**	**1.8**
	w	w	**-2.1**	**-3.5**	**1.5**
Classical Raking	d	d	-22.7	-30.6	-18.4
	d	w	-22.6	-30.5	-18.3
	w	d	**-1.7**	**-13.5**	**1.7**
	w	w	**-2.1**	**-13.7**	**1.3**
'ML' Raking	d	d	-22.8	-22.0	-18.4
	d	w	-22.7	-21.9	-18.3
	d	df	-22.3	-21.7	-17.9
	w	d	**-1.5**	**-2.7**	**1.9**
	w	w	**-2.1**	**-3.1**	**1.3**
	w	df	**-2.4**	**-3.3**	**1.1**
Additive nonresponse:					
GREG	d	d	-22.3	-21.8	-18.5
	d	w	-22.2	-21.7	-18.4
	w	d	**-1.6**	**-2.9**	**1.1**
	w	w	**-2.0**	**-3.1**	**0.8**
Classical Raking	d	d	-22.3	-30.2	-18.0
	d	w	-22.2	-30.1	-17.9
	w	d	**-1.5**	**-13.3**	**1.8**
	w	w	**-1.9**	**-13.5**	**1.4**
'ML' Raking	d	d	-22.4	-21.6	-18.0
	d	w	-22.3	-21.5	-17.9
	d	df	-22.0	-21.3	-17.6
	w	d	**-1.3**	**-2.4**	**2.0**
	w	w	**-1.9**	**-2.8**	**1.5**
	w	df	**-2.2**	**-3.0**	**1.3**

Figures in bold indicate the best approach under each nonresponse scenario

[1]see text following equation (4.3.11), where df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively

Table 4.6.5: Properties of linearization variance estimators when estimating total income from the SIE (R = 1000)

Estimator	w - or d - weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	d	d	10,338.8	-138.5 (6.9)	259.0	93.8
	d	w	10,339.2	-138.2 (6.9)	258.8	93.8
	w	d	10,377.9	**-99.5 (6.9)**	**240.0**	94.1
	w	w	10,376.8	**-100.5 (6.9)**	**240.3**	94.1
Classical Raking	d	d	10,338.8	-145.3 (6.9)	262.7	93.8
	d	w	10,339.2	-144.9 (6.9)	262.5	93.8
	w	d	10,370.0	**-106.1 (6.9)**	**243.1**	94.0
	w	w	10,376.9	**-107.2 (6.9)**	**243.5**	94.0
'ML' Raking	d	d	10,338.8	-152.7 (6.9)	266.9	93.9
	d	w	10,339.2	-152.4 (6.9)	266.7	93.9
	d	df	10,340.3	-151.3 (6.9)	266.1	94.0
	w	d	10,378.3	**-113.2 (6.9)**	**246.5**	94.0
	w	w	10,377.1	**-114.4 (6.9)**	**247.0**	94.0
	w	df	10,376.7	**-114.8 (6.9)**	**247.2**	94.0
Multiplicative nonresponse:						
GREG	d	d	8,104.7	-5,482.1 (7.4)	5,487.1	75.8
	d	w	8,105.5	-5,481.3 (7.4)	5,486.3	75.8
	w	d	13,214.5	**-372.3 (12.8)**	**549.7**	94.5
	w	w	13,210.9	**-375.9 (12.8)**	**551.7**	94.5
Classical Raking	d	d	8,104.7	-5,479.8 (7.4)	5,484.9	75.8
	d	w	8,105.5	-5,479.1 (7.4)	5,484.1	75.8
	w	d	13,214.1	**-370.4 (12.8)**	**549.4**	94.5
	w	w	13,210.4	**-374.2 (12.8)**	**551.5**	94.5
'ML' Raking	d	d	8,104.7	-5,478.1 (7.4)	5,483.1	75.8
	d	w	8,105.5	-5,477.3 (7.4)	5,482.3	75.8
	d	df	8,108.1	-5,474.7 (7.4)	5,479.7	75.9
	w	d	13,215.2	**-367.6 (12.9)**	**549.4**	94.5
	w	w	13,210.6	**-372.2 (12.9)**	**551.6**	94.5
	w	df	13,208.9	**-373.9 (12.9)**	**552.3**	94.5
Additive nonresponse:						
GREG	d	d	8,106.3	-5,508.5 (7.4)	5,513.5	75.6
	d	w	8,107.1	-5,507.7 (74)	5,512.7	75.6
	w	d	13,207.9	**-407.0 (12.8)**	**573.8**	94.3
	w	w	13,204.3	**-410.5 (12.8)**	**575.9**	94.3
Classical Raking	d	d	8,106.3	-5,506.6 (7.4)	5,511.6	75.7
	d	w	8,107.1	-5,505.9 (7.4)	5,510.9	75.7
	w	d	13,207.7	**-405.3 (12.8)**	**573.6**	94.1
	w	w	13,203.9	**-409.0 (12.8)**	**575.8**	94.1
'ML' Raking	d	d	8,106.3	-5,507.2 (7.4)	5,512.2	75.9
	d	w	8,107.1	-5,506.4 (7.4)	5,511.4	75.9
	d	df	8,109.7	-5,503.8 (7.4)	5,508.8	75.9
	w	d	13,208.9	**-404.6 (12.9)**	**574.8**	94.1
	w	w	13,204.2	**-409.2 (12.9)**	**577.3**	94.1
	w	df	13,202.5	**-411.0 (12.9)**	**578.1**	94.1

Figures in bold indicate the best approach under each scenario

[1]see text following equation (4.3.11), where choices df, d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively

[2] percentage of 95% normal-theory confidence intervals containing true value

Table 4.6.6: Percent relative bias of linearization variance estimators of expenditure and income totals from SIE (R= 1000)

Weighting Method	w - or d - weighted residuals	weight used for $\hat{B}$ in residual[1]	Percent Relative Bias of Standard Error Estimator	
			Expenditure	Income
Complete Response:				
GREG	d	d	0.7	-1.3
	d	w	0.7	-1.3
	w	d	1.3	-1.0
	w	w	1.3	-1.0
Classical Raking	d	d	0.7	-1.4
	d	w	0.7	-1.4
	w	d	1.2	-1.0
	w	w	1.2	-1.0
'ML' Raking	d	d	0.6	-1.5
	d	w	0.6	-1.5
	d	df	0.6	-1.4
	w	d	1.2	-1.1
	w	w	1.2	-1.1
	w	df	1.2	-1.1
Multiplicative nonresponse:				
GREG	d	d	-38.2	-40.4
	d	w	-38.2	-40.3
	w	d	**-0.3**	**-2.7**
	w	w	**-0.3**	**-2.8**
Classical Raking	d	d	-38.2	-40.3
	d	w	-38.2	-40.3
	w	d	**-0.3**	**-2.7**
	w	w	**-0.3**	**-2.8**
'ML' Raking	d	d	-38.2	-40.3
	d	w	-38.2	-40.3
	d	df	-38.2	-40.3
	w	d	**-0.3**	**-2.7**
	w	w	**-0.3**	**-2.7**
	w	df	**-0.4**	**-2.8**
Additive nonresponse:				
GREG	d	d	-38.1	-40.5
	d	w	-38.1	-40.5
	w	d	**-0.2**	**-3.0**
	w	w	**-0.2**	**-3.0**
Classical Raking	d	d	-38.1	-40.5
	d	w	-38.1	-40.5
	w	d	**-0.2**	**-3.0**
	w	w	**-0.2**	**-3.0**
'ML' Raking	d	d	-38.2	-40.5
	d	w	-38.2	-40.5
	d	df	-38.1	-40.4
	w	d	**-0.2**	**-3.0**
	w	w	**-0.3**	**-3.0**
	w	df	**-0.3**	**-3.0**

Figures in bold indicate the best approach under each nonresponse scenario

[1] see text following equation (4.3.11), where df , d and w correspond to $\hat{B}$ in (i), (ii) and (iii) respectively

Table 4.6.7 presents the properties of group jackknife estimators of the variance of the generalised regression point estimator of the total number of persons unemployed in the LFS study for different assumptions about nonresponse (the 'standard error estimate' in the table refers to the square root of the variance estimate). Since the jackknife calculations are very time-consuming, and larger biases of the standard error estimator compared to the linearization approach are observed (see Table 4.6.3 and 4.6.4), the jackknife method is not employed in this simulation study for the other generalised raking estimators. For the generalised regression calibration method, two versions of the grouped jackknife variance estimators are computed with number of groups $g = 38, 76$. In each case, the initial sample within each of the 19 strata is equally divided into $g/19$ random groups and deleted one at a time to create the replications.

Table 4.6.7: Properties of alternatives jackknife variance estimators of the GREG point estimator of the total unemployed from the LFS (R = 1000)

Nonresponse Model	g	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	Percent Relative Bias	RMSE of SE Estimator	Coverage[1] of Confidence Interval (%)
Complete	38	501.3	48.6 (4.1)	10.7	139.1	94.5
Response	76	508.8	56.0 (2.5)	12.4	97.6	95.9
Multiplicative	38	569.2	71.0 (3.8)	14.2	177.8	95.1
nonresponse	76	579.2	81.0 (4.2)	16.3	154.5	95.9
Additive	38	568.8	71.4 (5.4)	14.4	184.9	94.8
nonresponse	76	575.6	78.1 (4.1)	15.7	150.8	95.8

[1] percentage of 95% normal-theory confidence intervals containing true value

Table 4.6.7 shows evidence of positive bias, relative to the simulation standard error, for all grouped jackknife estimators. Slightly larger biases are observed for the estimators computed with more number of groups (i.e. $g = 76$) . However, it is important to observe that the RMSEs of these latter estimators are always smaller than those of the jackknife estimators computed with less number of groups. This indicates that standard errors of grouped jackknife variance estimators computed with more groups are smaller than those computed with fewer groups, offsetting the bias effect.

Under the complete response set-up, jackknife relative biases are always of larger magnitude than those observed from the linearization methods. The same applies under nonresponse for the linearization methods that use calibrated weights w_i to weight the

residuals (see Table 4.6.4). Some reduction in absolute bias is observed when jackknife is compared to linearization using d_i weighted residuals.

Valliant (2004) compared four versions of grouped jackknife variance estimators, with $g = 10, 25, 50, 100$, by conducting a simulation study using a poststratified population similar to the British LFS and selecting samples of size $n = 100, 250, \text{and } 500$. In line with the results presented in this section, he observed positive relative biases for all grouped jackknife variance estimates with biases ranging from 8.4% for the jackknife estimator with least number of groups to 20.4% for the jackknife estimator with most number of groups, for $n = 100$. He noted that the biases reduce for other sample sizes but the pattern of positive biases always persists.

Regarding confidence interval coverage, Table 4.6.7 shows at least 95% coverage for both group choices and alternative nonresponse assumption. The larger overestimation for the grouped jackknife estimator with $g = 76$ is accompanied by some overcoverage by confidence intervals, even though the excess above the nominal level is small. Valliant (2004) showed similar results for the two larger sample sizes with at least 95% coverage for all variance estimators.

The linearization estimators in this study give underestimates of variances and confidence intervals that tend to cover at lower, if small in some cases, than nominal level. On the other hand, the jackknife estimators tend to overestimate variances with estimated confidence intervals slightly above the nominal level.

4.7 Conclusions

The simulation study show little difference between the bias or variance properties of the three calibration estimators considered: the GREG estimator, the classical raking estimator and the maximum likelihood raking estimator. Some small differences in the distribution of extreme weights are observed. A few negative weights are observed for the GREG estimator, whereas weights are necessarily positive for both raking estimators. Some very large weights are observed for the maximum likelihood raking estimator.

Amongst the linearization variance estimators, the main finding is the contrast between the approach which weights residuals by the design weight and that which

weights them by the calibrated weight. It is found that the latter variance estimator tends always to have reduced bias and that this effect is very marked in the presence of nonresponse, when the former estimator could be severely negative biased. The bias of the latter estimator, if negative, is generally small and the coverage level of the associated confidence intervals is generally close to the nominal coverage. Alternative ways of weighting the observations in constructing the regression coefficients, when calculating the residuals in the linearization variance estimator, are considered but little effect is observed and there is no evidence that this choice is important in practice. In general, the findings for the categorical variables in the British Labour Force Survey are remarkably similar to the findings for the continuous variables in the German Income and Expenditure survey.

Unlike the linearization variance estimators, the simulation presented in this chapter shows that the grouped jackknife estimators of the variance of the generalised regression point estimator of the total number of persons unemployed in the LFS tend to be an overestimate. This overestimation causes some small overcoverage by confidence intervals. Thus, the jackknife approach results in nominal coverage but at the expense of larger overestimation. The jackknife method is much more intensive computationally than the linearization approach, but it does not require working out a variance expression for each particular parameter of interest, which might be a burden in some complex multipurpose surveys.

Chapter 5

Conclusions

This doctoral thesis focuses on understanding and dealing with unit nonresponse in sample surveys during and post data collection. The first part of this thesis (Chapter 2) relates to strategies that may be used prior and during data collection to enhance response rates. The following chapters (Chapter 3 and 4) refer to post-survey estimation methods to adjust and account for nonresponse. In this chapter the main findings of this doctoral research are summarised and some limitations and further work are discussed.

5.1 Summary and implications for survey practice

Chapter 2 illustrates the use of field process data or paradata, particularly interviewer call record and interviewer observation data, to separately model the process of establishing contact and cooperation with sample members in face-to-face surveys. It aims to better understand the process leading to contact or cooperation rather than focussing on predicting the final response outcome. It also introduces the reader to the analysis of call record data in a multilevel modelling framework, motivated by a range of both technical and substantive reasons. The analyses in this chapter use data from the Census Link Study 2001, which provides an exceptional opportunity to analyse the effectiveness of interviewer calling behaviours and strategies to establish contact and obtain subsequent cooperation, controlling for household and interviewer characteristics. The dataset combines rich paradata from six major UK interview administrated household surveys.

Results from this chapter indicate that time-varying call record information, such as features of the call history and of the current call, play a key role in predicting contact and the subsequent outcome of each call. For example, the results support earlier findings that weekday evenings and weekend daytime are, on average, the best times to establish contact with a household. Although, without a prior appointment, households

contacted at those times are more likely to refuse; the analysis here shows that they also have higher chances of making an appointment which might result in future cooperation. Of particular interest for survey agencies are interviewer strategies to achieve contact and gaining cooperation. The contact model shows some significant effects of such strategies, for example the probability of contact is higher at the next call if the interviewer left a card or message at a previous call. Regarding cooperation, characteristics of the doorstep interaction process between the interviewer and the householder, such as how contact was established and whether the householder asked questions or made comments, seem to be of relevance. This chapter provides substantial evidence that interviewer observations about a household and neighbourhood are useful for predicting the likelihood of contact and cooperation. Some of these observations are predictive of contact and cooperation before and after controlling for additional census information about the household. Interviewer observations, such as the presence of dependent children, type and condition of the house, might be regarded as proxies for census information that is usually unavailable. Area characteristics might also be considered as proxies for household characteristics and useful for predicting contact and cooperation. This research finds a number of significant effects of interviewer characteristics on the process leading to contact and cooperation. Important in explaining interviewer differences in contact rates are pay grade, qualifications and age. The attitude of the interviewer towards refusal conversion and the interviewer's self-confidence play an important role in the cooperation process. The length of interviewer experience, although not significant for achieving contact after controlling for other variables in the model, is significantly negatively associated with refusal at the doorstep. Some evidence for differential effects of fixed interviewer characteristics across the three non-participating outcomes -refusals, appointments made and other forms of postponement- is found. Unmeasured interviewer characteristics have a significant effect on contact and cooperation. However, the variation between interviewers in their cooperation rates is higher than the variation in their contact rates, providing some evidence that interviewer effects are more important for the process leading to cooperation. This might be due to the fact that this process depends much more on interviewer skills and behaviours and the interaction between the interviewer and the householder at the doorstep than the process leading to contact, which is more determined by timings and household characteristics. The influence of the interviewer random effect is the same across refusal, appointment made and other

forms of postponement. In contrast, this research shows evidence of differential effects of unmeasured household characteristics across the three non-participating outcomes. Household unobservables that are positively associated with refusal are negatively associated with appointments made and other forms of postponements.

The results in Chapter 2 have a number of potential implications for survey practice. The type of models presented and the variables identified as important to predict contact and cooperation may be used to inform the design of efficient and effective calling behaviours and follow-ups as well as responsive survey designs (Groves and Heeringa, 2006; Laflamme et al., 2008) - even in the absence of information like here from the census. For example, an interviewer or survey agency may be able to observe hints for a potential refusal early on, such as certain comments or questions from a householder or an increased number of initial or intermediate non-contacts, before a hard refusal occurs. Such hints may inform early intervention schemes that survey agencies can employ before the end of the data collection period to increase final response rates and to potentially reduce nonresponse bias. Survey organisations could respond to such difficult cases by changing the contacting strategy, for example, by offering a higher incentive or by sending a more experienced interviewer. A particular application of such models might be within the context of longitudinal surveys where call record data and a wide range of information on the sample member are available from previous waves. The models may also inform improvements for interviewer training and interviewer selection, for example, survey organisations may assess how to improve interviewers' training to best deal with the initial interaction with the householder at the doorstep. The research in this chapter highlights important advantages of gathering call record information and interviewer observations during data collection to inform the process leading to contact and cooperation. These variables could be used as proxies of household characteristics if census or administrative data are not available. This has also implications for survey agencies that need to carefully consider which types of paradata should be recorded at each call and how best to collect such data, including interviewer training. The significant interviewer effects in predicting contact and cooperation imply that survey organisations may be able to allocate certain interviewers to more difficult cases - at least within fieldwork constrains such as travelling and costs. The models developed in this study might also be useful to estimate response propensities to be used for adjustment and estimation at the data analysis stage, as investigated in the following chapter.

Chapter 3 explores alternative inverse probability weighted estimators for clustered nonresponse when cluster membership is observed for both responding and nonresponding units. That, for example, could be the case when the clusters are defined by interviewer workloads as in the previous chapter. This research considers three 'standard' ways of constructing inverse probability weights, including the use of multilevel models as in Chapter 2 and marginal models that ignore the clustering structure of the data. It also proposes a new approach using conditional maximum likelihood (CML). This chapter investigates to what extent inverse probability weights based on multilevel models result in more efficient estimates than those obtained by using simpler models that ignore the clustered data. A key aim is to construct weights which exploit the auxiliary information on cluster membership and other variables to correct for bias under cluster-specific non-ignorable (CSNI) nonresponse as proposed by Yuan and Little (2007), not just missing at random (MAR). It also examines variance estimators for each adjusted weighted estimator, assuming weights are treated as fixed. The properties of the alternative weighted estimators for two sampling designs and associated variance estimators are investigated through a simulation study. Results from an empirical application using data from the Expenditure and Food Survey 2001 are also presented.

The simulation study in Chapter 3 shows that under MAR nonresponse, when the cluster sizes are not large, the use of nonresponse weights based on predicted random effects can in fact bring negative relative bias in the inverse probability weighted estimator. The empirical application, however, does not show any disadvantages from the estimator based on predicted random effects compared to the others. If MAR is plausible, particularly for small cluster sample sizes, it seems reasonable to employ in practice simple response propensity weights based upon a marginal model for response rather than weights based on a multilevel model. On the other hand, under a CSNI nonresponse mechanism, not just MAR, the simulation results indicate that the marginal approach may be subject to bias. In this case, the new proposed approach using conditional maximum likelihood seems to perform the best and thus is is recommended. According to the simulation findings, the fixed effects estimator performs similarly to the CML estimator. This may indicate that in practice this estimator might often provide a reasonable approximation to the CML estimator, while requiring less computation time and not such strong model assumptions. Regarding the performance of the random effect estimator under CSNI, some potential benefits of this estimator over the

simple response propensity estimator based on a marginal model are observed, in particular for larger cluster sample sizes.

The simple variance estimators presented in Chapter 3, treating weights as fixed, show some large relative biases for the marginal and random effect estimators. It would be sensible to consider variance estimators that account for the nonresponse adjustments. Alternative variance estimation methods for a particular class of weighted estimators in the presence of nonresponse are discussed in Chapter 4.

Chapter 4 focuses on a particular class of weighting procedure called calibration. It reports a simulation study of the properties of three forms of generalized raking estimators: the GREG estimator, the classical raking estimator and the maximum likelihood raking estimator; and associated variance estimators with respect to the effects of both sampling and nonresponse. The simulation study is designed to mimic two major European surveys: the UK Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). The research in this chapter explores alternative forms of linearization variance estimators for generalized raking estimators in the presence of unit nonresponse. It also investigates one of the most frequently used replication methods, the jackknife method, of computing variances for complex sample surveys accounting for nonresponse.

The simulation study in Chapter 4 shows little difference between the bias or variance properties of the three calibration estimators considered. Some small differences in the distribution of extreme weights are observed: the maximum likelihood raking estimator has the most very large weights and the GREG estimator is the only one with a few negative weights. The main finding regarding the linearization variance estimators is the difference between the approach that weights the residuals by the design weight and the approach that weights them by the calibrated weight. The latter variance estimator tends to have smaller negative bias and this effect is very marked in the presence of nonresponse, when the estimator that weights the residuals by the design weight could be severely negative biased. Alternative ways of weighting the observations in constructing the regression coefficients, when calculating the residuals in the linearization variance estimator, are considered but little effect is observed and there is no evidence that this choice is important in practice. Regarding jackknife variance estimation, the simulation shows that the grouped jackknife estimators tend to be an overestimate. This overestimation causes some small overcoverage of confidence intervals.

5.2 Limitations and further work

A potential limitation of the study presented in Chapter 2 is that the available data are based on a non-random allocation of calling times to households. That is, the data are not obtained via a controlled experiment but reflect observational data. The models attempt to control for important household and interviewer characteristics likely to be associated with the interviewer decision on when to call. Nonetheless, as it is possible that the calling time may depend on unmeasured household and interviewer characteristics, the effects of calling times should be interpreted with caution and statements about possible casual effects should be limited. Another possible limitation of the data is that some information on specific interviewing strategies only reflects what an interviewer does in general (self-reported) and is not recorded at the call level (direct observation). For example, the variable indicating whether it is the interviewer's general practice to leave a card or message behind has no significant effect on contact; however, the time-varying covariate capturing the same information for each call is found significant. As suggested by Groves and Couper (1998), it may be preferable to ask interviewers to record their strategy for each call or household. More information at the call level may therefore be necessary to identify general trends on interviewer tailoring abilities.

The following are a number of specific recommendations for future research on paradata and nonresponse. The positive effect of the number of intermediate noncontact calls on refusal, discussed in Section 2.4.3, might provide some evidence to support the hypothesis that a noncontact call could in fact be a hidden evasion or refusal (Groves and Couper, 1998; Stoop, 2005). The lack of a correlation between the noncontact and refusal processes identified in earlier research (Lynn et al. 2002; Nicoletti and Perachi, 2005; Steele and Durrant, 2011) has so far not provided much support for this hypothesis. Further research might be needed to investigate this possible phenomenon, for example, including the additional outcome of a noncontact at a call in the modelling.

It may be argued that certain interviewers are better at gaining cooperation with harder cases. For example, more experienced interviewers may be more successful in dealing with householders that make negative comments or have questions. Effects of this type could help to inform the allocation of certain interviewers to potentially more

difficult households. Although the first part of this chapter explores interaction effects between interviewer characteristics and type of household in the context of contact, the section on the process leading to cooperation does not investigate such effects. Further work is needed to examine the hypothesis that some interviewers might be more successful in dealing with more difficult cases.

The overall aim of Chapter 2 is to contribute to a better understanding of the processes leading to contact and cooperation and the influence of factors that are associated with these processes. The results might inform strategies prior and during data collection to enhance response rates and to potentially reduce nonresponse bias. However, this research has not specifically investigated the relationship between nonresponse rates and nonresponse bias, which occurs when respondents differ from the nonrespondents with respect to the characteristics to be investigated. Further efforts are needed to investigate the use of paradata to reduce potential sources of nonresponse bias during the data collection process and to inform responsive survey designs with the aim of reducing such bias. Also, the potential uses of paradata in post-survey adjustments needs to be investigated further.

It is important to note that paradata can be subject to measurement errors and missing items and further research is needed to investigate the extent and potential sources of such error. Careful considerations need to be given on how to improve the quality of paradata as inaccurate information is likely to affect the resulting estimates and conclusions drawn from the application of such data.

Moving to Chapter 3, in line with the theory (Skinner and D'Arrigo, 2011) the empirical illustration suggests that under the missing at random assumption there is little to be gained from the method that account for the clustering in the data over the marginal approach. To gain further experience with these methods, further empirical studies could be conducted under the CSNI mechanism, not just MAR.

The conditional maximum likelihood approach shows a significant bias correction advantage under CSNI; however, it depends on the logistic form of the model in (3.2.3) and becomes increasingly computationally intensive as the sizes of the clusters grow. In addition, as observed in the simulation study, it can lead to more variable weights and can have efficiency disadvantages. A simpler modification of this approach would be to use what Little (1986) called response propensity stratification, forming classes by grouping values of the estimated CML weights and then replacing this weight by the inverse observed response rate in the group. This approach may be

less sensitive to the logistic link function assumption and may help smooth large values of estimated CML weights.

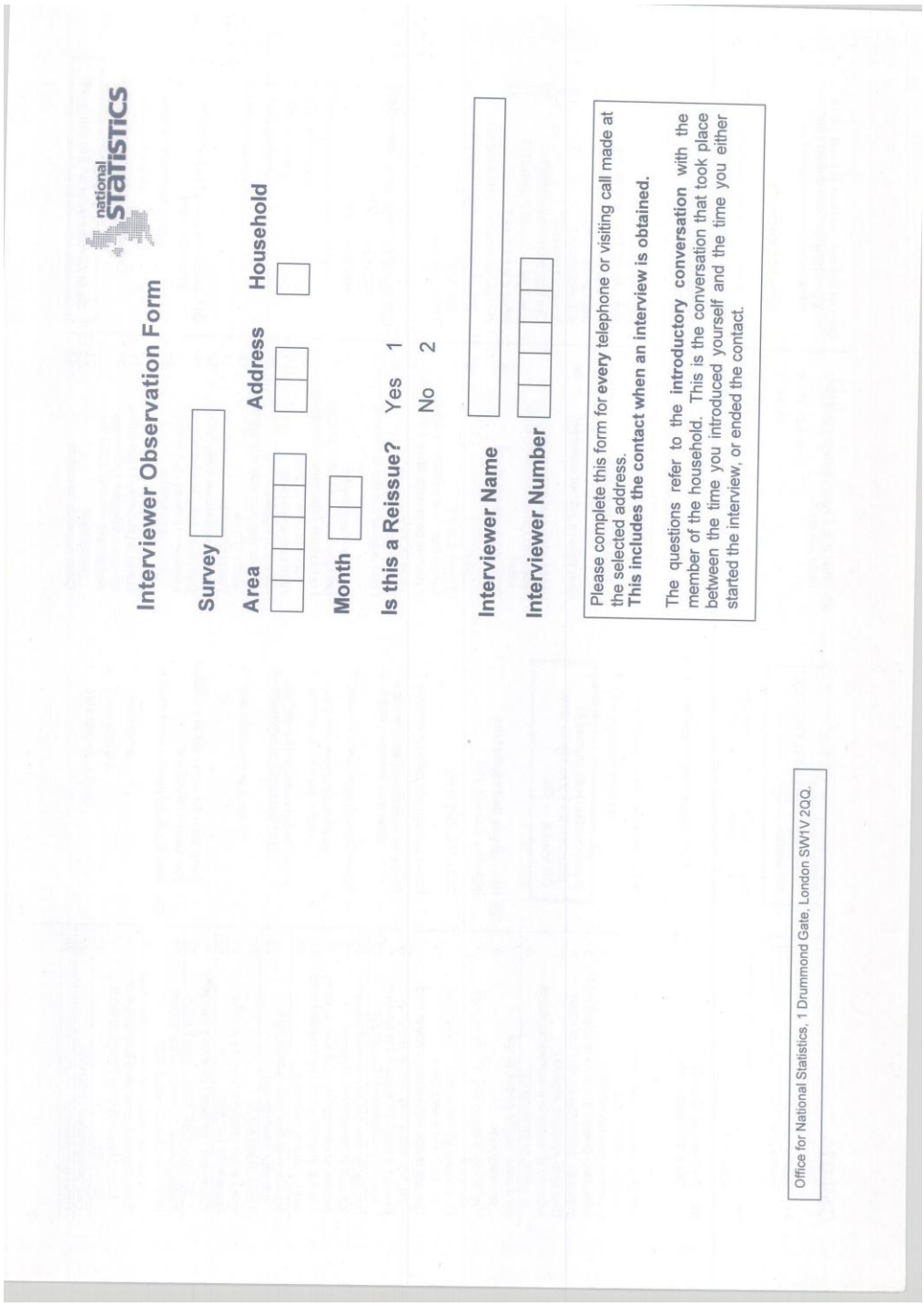
Chapter 3 presents an approximation of the required variance estimators treating the adjusted weights as fixed. Skinner and D'Arrigo (2011) outlined a more precise variance linearization approach that allows for variability on the estimated weights for the CML case. Further work might be needed to extend this approach to consider variance estimation that accounts for weighting adjustments for the other inverse probability weighted estimators presented in this chapter.

Finally, the Jackknife variance estimation methods presented in Chapter 4 require the calibration adjustments to be applied to each set of weights. For the classical raking ratio estimator and the 'maximum likelihood' raking estimator, this requires, in principle, iterating the raking method until convergence in each case. This imposes a high computational burden and results in the exclusion from this chapter of a jackknife variance estimator for these raking estimates. An alternative approach could be to reduce the number of iterations of the raking method, in particular by using a one-step jackknife (Shao and Tu, 1995, p.191). One version of the one-step jackknife, adopted by Canty and Davison (1999), is simply to stop after 'one step' of the raking adjustment to the initial replication weights, rather than continuing to convergence. This one step might consist of one step of Newton's method (Deville et al., 1993). Canty and Davison (1999) compared the performance of their one-step jackknife method with the jackknife method involving five iterations. They found that the performance of the variance estimator is actually worse for five iterations and concluded that "overall, the best jackknife strategy appears to be to use one iteration" (page 387). Alternatively, the one step of the raking adjustment could be applied to a set of calibrated weights formed by replacing the initial weights in (4.4.2) by the raked weights. These approaches should be asymptotically equivalent (Shao and Tu, 1995) but their finite sample properties require further study. Either of these one-step methods still requires the inversion of a matrix in the one step of Newton's method for each replicate and this could still be computationally heavy. This repeated inversion of a matrix could be avoided by use of the estimation function jackknife, studied by Rao and Tausi (2003). More research is needed to investigate alternative jackknife variance estimation methods.

Appendices

A1 - Interviewer Observation Form

The Interviewer Observation Form is a double-sided A5 booklet. Only the key pages are reproduced here; the pages for recording calls 02 onwards are broadly the same as the ones for recording call 01.



national STATISTICS

Interviewer Observation Form

Survey

Area Address Household

Month

Is this a Reissue? Yes 1
No 2

Interviewer Name

Interviewer Number

Please complete this form for **every** telephone or visiting call made at the selected address.
This includes the contact when an interview is obtained.

The questions refer to the **introductory conversation** with the member of the household. This is the conversation that took place between the time you introduced yourself and the time you either started the interview, or ended the contact.

Office for National Statistics, 1 Drummond Gate, London SW1V 2QQ.

Call 01

Q1. Date of call (ddmmyyyy)

Q2. Time of call (24h: hh.mm)

Q3. Outcome of this call
Code ALL that apply

HQ refusal	1
Ineligible: not built /demolished/derelict	2
Ineligible: vacant/empty	3
Ineligible: non-residential	4
Ineligible: residential but no resident hhld	5
Ineligible: communal estab / institution	6
Unable to establish eligibility	7
Address temporarily inaccessible due to weather or other causes	8
No answer on the phone or got answerphone	9
No face-to face contact with anyone at address	10
No contact with anyone at the address but spoke to a neighbour	11
Contact made with sampled household but not with a responsible resident	12
Contact made with responsible resident but not with selected respondent	13
Refusal at intro/bef int - by hhld member	14
Refusal at intro/bef int - by proxy	15
Refusal at intro/bef int – DK hhld or proxy	16
Refusal at intro/bef int - by selected resp	17
Refusal during interview	18
No interview due to language, age, infirmity, disability etc.	19
Appointment made	20
Interviewer withdrew to try again later	21
Appointment broken	22
Placement interview / checking /reminder call completed (diary surveys only)	23
Partial household interview completed	24
Hhld interview completed but non-contact with one or more elements	25
Hhld interview completed but refusal or incomplete interview by one or more elements	26
Full co-operation / interview	27
Other outcomes	28

If Q3 = 1 to 8, → Q11.
If non-contact (Q3 = 9 to 13), → Q4.
Otherwise, → Q5.

Q4. Did you leave a card or message at the address / on the phone / on the answerphone?

Card/message left at address	1
Message left on the phone	2
Message left on the answerphone	3
No card/message left	4

If non-contact with anyone at address (Q3 = 9 to 11), → Q10.
Otherwise, → Q5.

Q5. How did you first contact the household at this call?

Code ALL that apply

Spoke through entryphone/ intercom	1
Spoke through closed door/ window letter box/ door on a chain	2
Spoke through open door/ window	3
Spoke to informant who was outside the household unit	4
Went /invited inside the household unit and spoke to informant	5
Spoke over the telephone	6

Q6 to Q8 refer to the MAIN PERSON you made contact with.

Q6. Was the main person you talked to:

a man/boy?	1
a woman/girl?	2
Don't know, not sure	3

Q7. What is the approximate age of the main person?

Less than 16	1
16-34	2
35-59	3
60 and over	4
Don't know	5

Q8. Did the main person make any of the following comments during your introductory conversation?

Code ALL that apply

Main person did not comment	99
Positive/neutral comments:	
Received / remembered advance letter	1
Expecting someone to call	2
Make an appointment and come back	3
I'll think about it	4
Survey topic is important / or other positive comments about the survey topic	5
Enjoy doing surveys	6
Other positive /neutral comments	7
Negative comments:	
Not interested / can't be bothered	8
I'm too busy / Bad time / Just going out / About to go away	9
I don't know anything	10
We are not typical	11
Not capable / too sick/ old/ infirm	12
Waste of time	13
Waste of money	14
Government knows everything	15
Don't trust study is confidential	16
Invasion of privacy / too many personal questions	17
Don't trust surveys	18
Never do surveys / I hate forms	19
Already participated in surveys	20
Negative comments about the survey topic	21
Other negative comments	22

Q9. Did the main person ask any of the following questions during your introductory conversation?

Code ALL that apply

Main person did not ask questions	99
What is the purpose of the survey / What's it all about?	1
Who is paying for this/who is the sponsor?	2
What will happen to the information How will the results be used?	3
Why/how was I chosen?	4
How long will the interview take?	5
Who's going to see my answers?	6
Can I be identified?	7
Is it confidential?	8
Is this compulsory?	9
Can I get more information?	10
Can I get a copy of the results?	11
What's in it for me?	12
Do I get an incentive? How much is the incentive? When / how will I get paid?	13
Other questions	14

Q10. Diary surveys only. Non-diary surveys → Q11.

Was this call a:

call to make an appointment?	1
placement call?	2
checking call?	3
reminder call?	4
collection call?	5

Q11. When did you fill in the details about this call?

Immediately after the call	1
On the same day	2
On a different day from the call	3

Go to Accommodation Section (p.22)

Accommodation

Please complete for all addresses including ineligibles.

A1. What type of accommodation is it?

House or bungalow

Detached	1
Semi-detached	2
Terrace/end of terrace	3

Flat or maisonette

In a purpose built block	4
Part of a converted house / some other kind of building	5
Room or rooms	6
Caravan, mobile home or houseboat	7
Some other kind of accommodation	8
Don't know / not applicable / unable to code	9

If flat, maisonette or rooms (A1 = 4 to 6), → A2. Otherwise, → A4.

A2. Is the household's accommodation self-contained?

Yes, all rooms are behind a door that only this household can use	1
No	2
Don't know	3

A3. What is the floor level of this household's accommodation?

Basement or semi-basement	1
Ground floor (street level)	2
1st floor (floor above street level)	3
Second floor	4
Third floor	5
Fourth floor	6
Fifth to ninth floor	7
Tenth floor or higher	8
Don't know	9

A4. Are there any physical barriers to entry to the house/flat/building?

Code ALL that apply

Locked common entrance	1
Locked gates	2
Security staff or other gatekeeper	3
Entry phone access	4
None	5
Don't know	6

A5. Which of the following are visible at the sampled address?

Code ALL that apply

Burglar alarm	1
Security gate over front door	2
Bars/grills on any windows	3
Other security device(s) e.g. CCTV	4
Security staff or security lodge on estate or block	5
None of these	6
Don't know	7

Neighbourhood

The term "area" in the following questions refers to the area that you can see from the address.

N1.* Is the sampled house/flat/building in a better or worse condition than the others in the area?

Better	1
Worse	2
About the same	3
Unable to code	4

N2.* Are the houses/blocks in this area in a good or bad state of repair?

Mainly very good	1
Mainly good	2
Mainly fair	3
Mainly bad	4
Mainly very bad	5
Unable to code	6

N3.* How many boarded-up or uninhabitable buildings are there in this area?

None	1
One or two	2
A few	3
Several or many	4
Unable to code	5

N4.* Are most of the buildings in the area residential or commercial / non-residential?

All residential	1
Mainly residential with some commercial or non-residential	2
Mainly commercial or other non-residential	3
Unable to code	4

N5.* Is the house/flat part of a council or Housing Association housing estate?

Yes, part of a large council estate	1
Yes, part of a council block	2
No	3
Unable to code / not applicable	4

N6.* How safe would you feel walking alone in this area after dark?

Very safe	1
Fairly safe	2
A bit unsafe	3
Very unsafe	4

N7. Is this your final call to this household?

Yes 1 → ***Go to Household Section (p.24)***

No 2 → ***You are now finished until the next call***

Household information

H1. Interviewer code final outcome:

Full or partial interview / co-operation	1	→ *H2*
Refusal	2	
No interview due to language, age, infirmity etc	3	
Non-contact	4	→ *H8*
Ineligible	5	
HQ refusal	6	

H2. Please enter household details. Use "DK" for don't know and "Ref" for refused.

No. of adults (aged 16 or over) ☐ **No. of children** (less than 16) ☐

	Sex M - Male F - Female DK - Don't know Ref - Refused	**Age band** 1 16 - 34 2 35 - 59 3 60+ DK- Don't know Ref - Refused
Adult 1		
Adult 2		
Adult 3		
Adult 4		
Adult 5		
Adult 6		
Adult 7		

H3. Are there any children aged 5 or under?

Yes, definitely	1
Possibly	2
No	3
Don't know	4
Refused	5

H4. Is any adult in paid work?

Yes	1
No	2
Don't know /Refused	3

H5. How long has the household lived at this address?

Months → ☐

OR

Years → ☐

Don't know/ refused 99

H6. Do you know or think the occupants are:

Code from observation
Code ALL that apply

White	1
Mixed	2
Asian (Indian, Pakistani, Bangladeshi, other)	3
Black (Caribbean, African, other)	4
Chinese and other ethnic group	5
Don't know	8

National Travel Survey → H7.
Other surveys → H8.

H7. Does the household at present own or have available for use any cars or vans?

Include any company cars or vans if available for private use.

Yes	1
No	2
Don't know /refused	3

H8. Please check that you have completed the Accommodation & Neighbourhood Section (p.22-23) and enter the number of calls made to this address → ☐

You have finished at this address.
Thank you for completing the questionnaire.
Please key the information into your laptop at home and return the questionnaires to the office when you have completed this quota.

A2 - R code to compute weighted estimates of totals for the M, FE and RE weighting methods and CNI1 mechanism.

```
# This script run simulation study for population of 200 clusters of 10 units each, where 50 clusters are selected and all elements within clusters are sampled.

# Population 3: delta=5 gamma1=0 gamma0=1

# Overall response rate 70%.

library(VGAM)
library(MASS)

pop <- read.table("C:\\...\\population3.dat", header = TRUE)

M<-200                  # number of clusters in the population
Ni<-rep(10,200)         # cluster sizes
N<-sum(Ni)              # population size
w.parameter<-1          # standard deviation w2=1
w2.parameter<-1         # w2=1
thau.parameter<-1       # standard deviation thau2=1
m<-50                   # PSU sample size (clusters)
n<-10                   # SSU sample size (households)

r<-1000
true.ypop.Total<-sum(pop$yij)
yPSW.Point.Est.Total.logit<-array(rep(0,r),dim=c(1,1,r))
yFIXED.Point.Est.Total.logit<-array(rep(0,r),dim=c(1,1,r))
yRMC1.Point.Est.Total<-array(rep(0,r),dim=c(1,1,r))
Tyreg.PSW<-array(rep(0,r),dim=c(1,1,r))
Tyreg.FIXED<-array(rep(0,r),dim=c(1,1,r))
Tyreg.RMC1<-array(rep(0,r),dim=c(1,1,r))
overall.response<-array(rep(0,r),dim=c(1,1,r))
inclusion.probi<-m/M    # constant inclusion prob within clusters
inc.prob.ij<-rep(inclusion.probi,N)
pop<-cbind(pop,inc.prob.ij)

# SIMULATION LOOP
set.seed(38)
seeds.in.r <- sample(c(0:2023), size=r, replace=F)
for (j in 1:r){
cat(date(),"starting simulation loop pop1, r=",j,"\n")
set.seed(seeds.in.r[j])
# DRAW THE SAMPLE
# Sample m=50 PSU from population
PSU.id<-sample(1:M, m, F)          # select 50 PSU (clusters)
PSU.id<-sort(PSU.id)
# Select all elements from each PSU
z <- 0
while(z<20) {
sample.n<-NA  # initialize sample matrix
for (h in PSU.id) {
sample<-pop[pop$i==h,]
sample.n<-rbind(sample.n,sample)
}
sample.n<-sample.n[-c(1),]
sample.size<-dim(sample.n)[1]      # total sample size
sample.n$uij<-runif(sample.size,0,1)
for (k in 1:sample.size){
```

```
if(sample.n$uij[k]<=sample.n$Prij[k]) sample.n$rij[k]=1
if(sample.n$uij[k]>sample.n$Prij[k]) sample.n$rij[k]=0
}
sample.r<-sample.n[sample.n$rij==1,]          # delete non-respondents
sample.size.r<-dim(sample.r)[1]               # sample size respondents
ri<-NA
yi<-NA
for(k in PSU.id){
        ri.loop<-length(sample.n[sample.n$i==k & sample.n$rij==1,1])
        yi.mean<-mean(sample.n[sample.n$i==k & sample.n$rij==1,6])
        ri<-rbind(ri,ri.loop)
        yi<-rbind(yi,yi.mean)
        }
ri<-ri[-c(1),]
yi<-yi[-c(1),]
if (length(yi[!is.nan(yi)])==m) { z <- 20 } else { z <- z+1 }
}
sample.r$response.ratei<-rep(ri/n,ri)
sample.r$wij<-(sample.r$inc.prob.ij*sample.r$response.ratei)^(-1)
overall.response[j]<-sample.size.r/(m*n)

# M estimator (using the link 'logit')
Propensity.model.logit<-glm(rij ~ x1ij,family = binomial(link="logit"), data = sample.n)
phi.ij.logit<-
exp(coef(Propensity.model.logit)[1]+coef(Propensity.model.logit)[2]*sample.r$x1ij))/(1+(exp(coef(Propen
sity.model.logit)[1]+coef(Propensity.model.logit)[2]*sample.r$x1ij)))
weight.ij.logit<-(inclusion.probi*phi.ij.logit)^(-1)
yPSW.Point.Est.Total.logit[j]<-sum(weight.ij.logit*sample.r$yij)

lambda1<-(sum((sample.r$inc.prob.ij*phi.ij.logit)^(-1)*t(sample.r$x1ij)*sample.r$x1ij)^(-
1))*(sum((sample.r$inc.prob.ij*phi.ij.logit)^(-1)*t(sample.r$x1ij)*sample.r$yij))
Txs<-sum(((sample.n$inc.prob.ij)^(-1))*sample.n$x1ij)
Tx1<-sum(((sample.r$inc.prob.ij*phi.ij.logit)^(-1))*sample.r$x1ij)
Tyreg.PSW[j]<-yPSW.Point.Est.Total.logit[j]+((Txs-Tx1)*lambda1)          # GREG using M weights

# FE estimator (using the link 'logit')
sample.n$uj<-as.factor(sample.n$i)
sample.r$uj<-as.factor(sample.r$i)

Fixed.model.logit<-glm(rij ~ x1ij + uj, family = binomial(link="logit"), data = sample.n)
uj.logit<-
c(0,coef(Fixed.model.logit)[3],coef(Fixed.model.logit)[4],coef(Fixed.model.logit)[5],coef(Fixed.model.logit)
[6],coef(Fixed.model.logit)[7],coef(Fixed.model.logit)[8],coef(Fixed.model.logit)[9],coef(Fixed.model.logit)
[10],coef(Fixed.model.logit)[11],coef(Fixed.model.logit)[12],coef(Fixed.model.logit)[13],coef(Fixed.model.l
ogit)[14],coef(Fixed.model.logit)[15],coef(Fixed.model.logit)[16],coef(Fixed.model.logit)[17],coef(Fixed.m
odel.logit)[18],coef(Fixed.model.logit)[19],coef(Fixed.model.logit)[20],coef(Fixed.model.logit)[21],coef(Fix
ed.model.logit)[22],coef(Fixed.model.logit)[23],coef(Fixed.model.logit)[24],coef(Fixed.model.logit)[25],coe
f(Fixed.model.logit)[26],coef(Fixed.model.logit)[27],coef(Fixed.model.logit)[28],coef(Fixed.model.logit)[29
],coef(Fixed.model.logit)[30],coef(Fixed.model.logit)[31],coef(Fixed.model.logit)[32],coef(Fixed.model.logi
t)[33],coef(Fixed.model.logit)[34],coef(Fixed.model.logit)[35],coef(Fixed.model.logit)[36],coef(Fixed.mode
l.logit)[37],coef(Fixed.model.logit)[38],coef(Fixed.model.logit)[39],coef(Fixed.model.logit)[40],coef(Fixed.
model.logit)[41],coef(Fixed.model.logit)[42],coef(Fixed.model.logit)[43],coef(Fixed.model.logit)[44],coef(F
ixed.model.logit)[45],coef(Fixed.model.logit)[46],coef(Fixed.model.logit)[47],coef(Fixed.model.logit)[48],c
oef(Fixed.model.logit)[49],coef(Fixed.model.logit)[50],coef(Fixed.model.logit)[51])

u.ij.logit<-rep(uj.logit,ri)    # create fixed effects for each respondent
f.phi.ij.logit<-
(exp(coef(Fixed.model.logit)[1]+coef(Fixed.model.logit)[2]*sample.r$x1ij+u.ij.logit))/(1+(exp(coef(Fixed.
model.logit)[1]+coef(Fixed.model.logit)[2]*sample.r$x1ij+u.ij.logit)))
f.weight.ij.logit<-(inclusion.probi*f.phi.ij.logit)^(-1)
```

```
yFIXED.Point.Est.Total.logit[j]<-sum(f.weight.ij.logit*sample.r$yij)

lambda2<-(sum((sample.r$inc.prob.ij*f.phi.ij.logit)^(-1)*t(sample.r$x1ij)*sample.r$x1ij)^(-
1))*(sum((sample.r$inc.prob.ij*f.phi.ij.logit)^(-1)*t(sample.r$x1ij)*sample.r$yij))
Tx2<-sum(((sample.r$inc.prob.ij*f.phi.ij.logit)^(-1))*sample.r$x1ij)
Tyreg.FIXED[j]<-yFIXED.Point.Est.Total.logit[j]+((Txs-Tx2)*lambda2)          # GREG using FE
weights

# RE estimator
Response.model.RMC1<-glmmPQL(rij ~ x1ij, random = ~1|i, family = binomial(link="logit"), data =
sample.n)
ranef.ij<-rep(ranef(Response.model.RMC1,drop = TRUE)[[1]],ri)    # create random effects for each
respondent
phi.ij.RMC1<-
(exp(fixef(Response.model.RMC1)[1]+fixef(Response.model.RMC1)[2]*sample.r$x1ij+ranef.ij))/(1+(exp(
fixef(Response.model.RMC1)[1]+fixef(Response.model.RMC1)[2]*sample.r$x1ij+ranef.ij)))
weight.ij.RMC1<-(inclusion.probi*phi.ij.RMC1)^(-1)
yRMC1.Point.Est.Total[j]<-sum(weight.ij.RMC1*sample.r$yij)

lambda3<-(sum((sample.r$inc.prob.ij*phi.ij.RMC1)^(-1)*t(sample.r$x1ij)*sample.r$x1ij)^(-
1))*(sum((sample.r$inc.prob.ij*phi.ij.RMC1)^(-1)*t(sample.r$x1ij)*sample.r$yij))
Tx3<-sum(((sample.r$inc.prob.ij*phi.ij.RMC1)^(-1))*sample.r$x1ij)
Tyreg.RMC1[j]<-yRMC1.Point.Est.Total[j]+((Txs-Tx3)*lambda3)          # GREG using RE weights
}
```

A3 - Area of residence

Area	*Counties*
1	Cleveland, Cumbria and Durham
2	Northumberland, Tyne & Wear and Humberside
3	North Yorkshire, West Yorkshire and South Yorkshire
4	Derbyshire, Leicestershire and Nottinghamshire
5	Lincolnshire, Northamptonshire and Cambridgeshire
6	Norfolk and Suffolk
7	Bedfordshire, Hertfordshire and Essex
8	Inner London and Outer London
9	East Sussex, West Sussex, Kent and Surrey
10	Hampshire, Isle of Wight, Dorset and Wiltshire
11	Berkshire, Buckinghamshire and Oxfordshire
12	Avon, Gloucestershire and Somerset
13	Cornwall and Devon
14	Hereford & Worcester, Shropshire and Staffordshire
15	Warwickshire and West Midlands
16	Cheshire and Merseyside
17	Greater Manchester and Lancashire
18	Clwyd, Dyfed, Gwent and Gwynedd
19	Mid Glamorgan, South Glamorgan, West Glamorgan and Powys
20	Border, central and Dumfries & Galloway
21	Fife and Grampian
22	Highland, Tayside and Northern & Western Isles
23	Lothian and Strathclyde

A4 - R code to compute calibrated weights using linear function

```
# FUNCTION CALIBRATION
# This function compute the calibration weights according to Deville and Sarndal (1992) method for
different distance functions.

Var.Margin.1 <- Sample$STR1
Var.Margin.2 <- Sample$STR2
Var.Margin.3 <- Sample$STR3

# FIND THE LABEL USED FOR THE FIRST MARGIN
Table <- table(Var.Margin.1)
Modalities.Margin.1 <- dimnames(Table)[[1]]
mode(Modalities.Margin.1) <- "numeric"
NB.Modalities.Margin.1 <- length(Modalities.Margin.1)

# FIND THE LABEL USED FOR THE SECOND MARGIN
Table <- table(Var.Margin.2)
Modalities.Margin.2 <- dimnames(Table)[[1]]
mode(Modalities.Margin.2) <- "numeric"
NB.Modalities.Margin.2 <- length(Modalities.Margin.2)

# FIND THE LABEL USED FOR THE THIRD MARGIN
Table <- table(Var.Margin.3)
Modalities.Margin.3 <- dimnames(Table)[[1]]
mode(Modalities.Margin.3) <- "numeric"
NB.Modalities.Margin.3 <- length(Modalities.Margin.3)

# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE FIRST MARGIN
Mat.Margin.1 <- matrix(rep(0,times=NB.Modalities.Margin.1*Sample.Size),ncol=NB.Modalities.Margin.1,
nrow=Sample.Size)
for(j in (1:NB.Modalities.Margin.1))
        {
        Modalities <- Modalities.Margin.1[j]
        Mat.Margin.1[,j] <- as.numeric(Var.Margin.1 == Modalities)
        }

# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE SECOND MARGIN
Mat.Margin.2 <- matrix(rep(0,times=NB.Modalities.Margin.2*Sample.Size),ncol=NB.Modalities.Margin.2,
nrow=Sample.Size)
for(j in (1:NB.Modalities.Margin.2))
        {
        Modalities <- Modalities.Margin.2[j]
        Mat.Margin.2[,j] <- as.numeric(Var.Margin.2 == Modalities)
        }

# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE THIRD MARGIN
Mat.Margin.3 <- matrix(rep(0,times=NB.Modalities.Margin.3*Sample.Size),ncol=NB.Modalities.Margin.3,
nrow=Sample.Size)
for(j in (1:NB.Modalities.Margin.3))
        {
        Modalities <- Modalities.Margin.3[j]
        Mat.Margin.3[,j] <- as.numeric(Var.Margin.3 == Modalities)
        }
Calibration.Matrix.X <- cbind(Mat.Margin.1,Mat.Margin.2,Mat.Margin.3)          # Sample value of the
calibration variables

Tot <- c(3372,4568,8031,5611,3722,2943,5956,11712,7420,6166,3960,3905,2814,4406,6030,4411,
```

```
7314,2927,2710,1118,1841,1334,6291,930,874,820,809,724,711,710,664,720,44624,869,849,828,773,696,
689,805,744,816,49907,2994,3678,3297,2460,719,3359,4775,3744,2729,923,2301,3005,2781,2075,731,2605,
3337,2971,2307,795,3117,4011,3363,2806,1251,3651,5022,4025,2956,1426,2509,3186,2826,2251,1047,272
9,3668,3155,2706,1271)                  # Vector of known population margins

# FIND THE LABEL USED FOR THE SURVEY VARIABLE
Table <- table(Sample$EMP)
Modalities.EMP <- dimnames(Table)[[1]]
mode(Modalities.EMP) <- "numeric"
NB.Modalities.EMP <- length(Modalities.EMP)

# CREATE THE MATRIX OF AUXILIARY VARIABLES FOR THE SURVEY VARIABLE
Mat.EMP <-
matrix(rep(0,times=NB.Modalities.EMP*Sample.Size),ncol=NB.Modalities.EMP ,nrow=Sample.Size)
for(j in (1:NB.Modalities.EMP))
        {
        Modalities <- Modalities.EMP[j]
        Mat.EMP[,j] <- as.numeric(Sample$EMP == Modalities)
        }

"CALIBRATION"<-
function(Distance.Function.Number, Mat.X.s, Pop.Total.X, Vect.Pi.s, L, U)
{
        Sample.Size <- length(Vect.Pi.s)
        Design.Weights <- 1/Vect.Pi.s
        f <- rep(1, times = Sample.Size)
        Mat.X.s <- as.matrix(Mat.X.s)
        Pop.Total.X <- as.vector(Pop.Total.X)
        if(Distance.Function.Number == 1) {
                Calibration.Weights.s <- CalibDeville.f(F1.f, F1der.f,
                        Mat.X.s, Pop.Total.X, Design.Weights, f)
        }
        if(Distance.Function.Number == 2) {
                Calibration.Weights.s <- CalibDeville.f(F2.f, F2der.f,
                        Mat.X.s, Pop.Total.X, Design.Weights, f)
        }
        if(Distance.Function.Number == 4) {
                Calibration.Weights.s <- CalibDeville.f(F4.f, F4der.f,
                        Mat.X.s, Pop.Total.X, Design.Weights, f)
        }
                GINVERSE.T.Mat.X.s.Diag.C.s <- GINVERSE(T.Mat.X.s.Diag.C.s %*%
                        Mat.X.s)
                T.Pop.Total.X.Minus.Est.Total.X <- t(Pop.Total.X - Est.Total.X)
                Calibration.Weights.s <- T.Pop.Total.X.Minus.Est.Total.X %*%
                        GINVERSE.T.Mat.X.s.Diag.C.s %*% T.Mat.X.s.Diag.C.s
                Calibration.Weights.s <- Calibration.Weights.s * Vect.Pi.s
                Calibration.Weights.s <- T.Vect.1.s + Calibration.Weights.s
                Calibration.Weights.s <- as.vector(Calibration.Weights.s)
        }
# OUTPUT
        Calibration.Weights.s
}
"CalibDeville.f"<-
function(F.func, Fder.func, X, Tx, d, f, limit = 10, eps = 1e-005, ets =
        1e-005)
{
# Computes weights using several alternative distance funtions
# proposed by Deville & Särndal(1992).
# Input parameters:
#  F.func - function defining calibration distance from Deville & Sarndal;
```

```
#  Fder.func - gradient of function defining calibration distance;
#  X      - sample data matrix for auxiliary (x) variables to be used for calibration;
#  Tx     - vector of population totals for calibration;
#  d      - vector of design weights for each sample unit;
#  f      - vector for scale factors; currently used only for vector of 1s;
#  limit  - maximum number of iterations to perform if convergence not achieved.
# Output produced:
#  vector of calibrated weights;
# Residual function that defines nonlinear system to be solved to obtain lambda:
          residuos <- function(lamb, F.func, X, Tx, d, f)
          {
# Function for computing residuals of linear model given X
# Converts data into proper object classes
                    lamb <- matrix(lamb, ncol = 1)
                    d <- matrix(d, ncol = 1)
                    f <- matrix(f, ncol = 1)
                    Tx <- matrix(Tx, ncol = 1) # Compute required residuals
                    u <- X %*% lamb
          # Compute estimated total of x variables using HT estimator
                    Txpi <- t(X) %*% d
                    t(X) %*% ((F.func(u * f) - 1) * d) - (Tx - Txpi)
          }
# Function to compute Jacobian of specified distance function
# needed for solving for lambda
          jacobiano <- function(lamb, Fder.func, X, Tx, d, f)
          {
# Converts data into proper object classes
                    lamb <- matrix(lamb, ncol = 1)
                    d <- matrix(d, ncol = 1)
                    f <- matrix(f, ncol = 1)
                    Tx <- matrix(Tx, ncol = 1)
                    u <- X %*% lamb
                    t(X) %*% diag(as.vector(d * (Fder.func(f * u) * f)), nrow =
                              length(d)) %*% X
          }
# Compute estimated total of x variables using HT estimator
          Txpi <- t(X) %*% d          # Initializing lambda
          lamb <- GINVERSE(t(X) %*% diag(d) %*% X) %*% (Tx - Txpi)
          u <- X %*% lamb# Calibration weights
          h <- max(u)
          if (h > 0.99) {
                    position <- cbind(u,c(1:length(u)))
                    x.value <- X[position[u==h,2],]
                    tita <- 0.99/(x.value %*% lamb)
                    lamb <- c(tita) * lamb
          }
          # Initializing values required for solution
          Func <- residuos(lamb, F.func, X, Tx, d, f)
          Jota <- jacobiano(lamb, Fder.func, X, Tx, d, f)
          delta <- GINVERSE(Jota) %*% ( - Func)
          it <- 1    # Computes lambda by Newton's method
          while(((sum(delta^2) >= eps || sum(abs(Func)) >= ets) && (it <- it + 1) <
                    limit)) {
                    Func <- residuos(lamb, F.func, X, Tx, d, f)
                    Jota <- jacobiano(lamb, Fder.func, X, Tx, d, f)
                    delta <- GINVERSE(Jota) %*% ( - Func)
                    lamb <- lamb + delta
                    u <- X %*% lamb# Calibration weights
                    h <- max(u)
                    if (h > 0.99) {
```

```
                        lamb.before <- lamb - delta
                        position <- cbind(u,c(1:length(u)))
                        x.value <- X[position[u==h,2],]
                        tita <- (0.99 - x.value %*% lamb.before)/(x.value%*%(lamb - lamb.before))
                        lamb <- lamb.before + c(tita) * (lamb - lamb.before)
                        }
        }
                u <- X %*% lamb# Calibration weights
                w.v <- as.vector(d * F.func(f * u))

        # Defines output to be provided by function
        return(w.v)
}
"F1.f"<- function(u) {1 + u }
"F2.f"<- function(u) {exp(u)}
"F4.f"<- function(u) {(1 - u)^(-1)}
"F1der.f"<- function(u){rep(1, length(u))}
"F2der.f"<- function(u) {exp(u)}
"F4der.f"<- function(u) {(1 - u)^(-2)}

"GINVERSE"<- function(x, tol = sqrt(.Machine$double.eps)) {
        if(length(dim(x)) > 2)
                stop("x must be a matrix or vector")
        svdX <- svd(x)
        if(is.complex(x))
                svdX$u <- Conj(svdX$u)
        NotZero <- svdX$d > tol * svdX$d[1]
        ans <- if(all(NotZero)) svdX$v %*% ((1/svdX$d) * t(svdX$u)) else if(!
                any(NotZero)) {
                if(is.matrix(x))
                array(0, dim(x)[2:1])
                else matrix(0, 1, length(x))
        }
        else svdX$v[, NotZero] %*% ((1/svdX$d[NotZero]) * t(svdX$u[, NotZero]))
        attr(ans, "rank") <- sum(NotZero)
        ans
}

Weights.Cal <- CALIBRATION(1, Calibration.Matrix.X, Tot, Sample$Pi, 0.62, 1.85)          # Vector of
calibrated weights using the linear function
```

A5 - R code to compute weighted residuals

```
# FUNCTION CALIBRATION.RESIDUALS
#  * The "Weights.of.Residuals" and the "Weights.for.Coef.Regression"
#    can be (i) the initial weights of function "CALIBRATION.WEIGHTS"
#    or (ii) the final weights of function "CALIBRATION.WEIGHTS".
#  * The weights for "Weights.for.Coef.Regression" must be positive.

CALIBRATION.RESIDUALS <-
function(Vect.Y,Weights.of.Residuals,Weights.for.Coef.Regression,Mat.Cal.Var)
        {
# PUT THE DATA IN A SINGLE DATA FRAME

Data <- data.frame(cbind(Vect.Y,Mat.Cal.Var))

# CREATE THE NAME AND THE FORMULA FOR THE WEIGHTED LEAST SQUARE FIT
```

```
NB.Cal.Var <- as.numeric(dim(Mat.Cal.Var)[2])
List.Name.Variables <- c("Y")
Formula <- "Y ~ -1"

for(i in (1:(NB.Cal.Var)))
        {
        Name.New.Variable <- paste("X",i,sep="")
        Formula <- paste(Formula," + ",Name.New.Variable,sep="")
        List.Name.Variables <- c(List.Name.Variables,Name.New.Variable)
        }

Formula <- as.formula(Formula)
List.Blank <- dimnames(Data)[[1]]
Names.List <- list(List.Blank,List.Name.Variables)
dimnames(Data) <- Names.List

# THE WEIGHTED LEAST SQUARE FIT

Model.Fit <- lm(formula=Formula,data=Data,weights=Weights.for.Coef.Regression)                #
One option for singular is put: singular.ok=T

# THE WEIGHTED RESIDUALS

Residuals <- residuals(Model.Fit) * Weights.of.Residuals

# OUTPUT

Residuals
        }
```

References

Agresti, A. (2002). Categorical Data Analysis. 2nd Ed. Hoboken: Wiley.

Bartholomew, D. (1961). A Method of Allowing for Not-at-home Bias in Sample Surveys. *Applied Statistics,* **10**, 52-59.

Bates, N., Dahlhamer, J. and Singer, E. (2008). Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse. *Journal of Official Statistics*, **24**, 4, 591-612.

Beerten, R. and Freeth, S. (2004). Exploring Survey Nonresponse in the UK: The Census-Survey Nonresponse Link Study. Office for National Statistics, *Working Paper*, 1-16.

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics,* **4**, 25, 1-260.

Binder, D.A. and Théberge, A. (1988). Estimating the Variance of Raking Ratio Estimators. *Canadian Journal of Statistics*, **16**, 47-55.

Blom, A.G. and Blohm, M. (2007). The Effects of First Contact by Phone: Evidence for the European Social Survey. Paper presented at the *18th International Workshop on Survey Household Nonresponse*, Southampton, UK.

Blom, A.G., de Leeuw, E.D. and Hox, J.J. (2010). Interviewer Effects on Nonresponse in the European Social Survey. Institute for Social and Economic Research, *Working Paper Series 2010-25*.

Brackstone, G. J. and Rao, J. N. K. (1979). An Investigation of Raking Ratio Estimators. *Sankhyã*, *Series C*, **41**, 97-114.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Belmont: Wadsworth.

Brick, J.M. and Allen, B. and Cunningham, P. (1996). Outcomes of a Calling Protocol in a Telephone Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 142-149.

Browne, W.J. (2009). *MCMC Estimation in MLwiN v2.10*. Centre for Multilevel Modelling, University of Bristol.

Canty, A. J. and Davison, A. C. (1999). Resampling-based Variance Estimation for Labour Force Surveys. *The Statistician*, **48**, 379–391.

Cao, W., Tsiatis, A.A. and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika,* **96**, 723-734.

Carrell, S.E., and West, J.E. (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, **118**, 3, 409-32.

Cassel, C.M., Särndal, C-E. and Wretman, J.H. (1983). Some Use of Statistical Models in Connection with the Nonresponse Problem. In *Incomplete Data in Sample Surveys III, Symposium on Incomplete Data, Proceedings*, Ed. W.G. Madow and I. Olkin, 143-160. New York: Academic Press.

Chang, T. and Kott, P.S. (2008) Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. *Biometrika*, **95**, 555-571.

Cobben, F. and Schouten, B. (2007). A Follow-up with Basic Questions of Nonrespondents to the Dutch Labour Force Survey. *Discussion paper 07011, Statistics Netherlands.*

Couper, M.P. (1998). Measuring Survey Quality in a CASIC Environment. *Proceedings of the Survey Research Methods Section, American Statistical Association*, **48**, 743-772.

Couper, M.P. (2001). Web Surveys: a Review of Issues and Approaches. *Public Opinion Quarterly*, **64**, 464-94.

Couper, M.P. and Groves R.M. (1992). The Role of the Interviewer in Survey Participation, *Survey Methodology*, **18**, 263-278.

Cunningham, P., Martin, D. and Brick, M. (2003). An Experiment in Call Scheduling. *Proceedings of the Section on Survey Research Methods, American Association for Public Opinion Research*, 59-66.

Curtin, R., Presser, S., Singer, E. (2005). Changes in Telephone Survey Nonresponse Over the Past Quarter Century. *Public Opinion Quarterly*, **69**, 1, 87-98.

Da Silva, D.N. and Opsomer, J.D. (2004). Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism. *Survey Methodology,* **30**, 45-55.

Da Silva, D.N. and Opsomer, J.D. (2006). A Kernel Smoothing Method to Adjust for Unit Nonresponse in Sample Surveys. *Canadian Journal of Statistics,* **34***,* 563-579.

de Heer, W. (1999). International Response Trends: Results of an International Survey. *Journal of Official Statistics*, **15**, 2, 129-142.

de Leeuw E., de Heer W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In: *Survey Nonresponse*, Groves Robert M., Dillman Don A., Eltinge John L., Little Roderick J. A. New York: Wiley, 41-54.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, **11**, 427-444.

Demnati, A. and Rao, J.N.K. (2004). Linearization Variance Estimators for Survey Data (with discussion). *Survey Methodology*, **30***,* 17-34.

Deville, J-C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. *Survey Methodology*, **25**, 193-203.

Deville, J-C. and Särndal, C-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 376-82.

Deville, J-C., Särndal, C-E. and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, **88**, 1013-20.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Ed. Oxford: Oxford University Press.

Durrant, G.B. and Steele, F (2009). Multilevel Modelling of Refusal and Noncontact Nonresponse in Household Surveys: Evidence from Six UK Government Surveys. *Journal of the Royal Statistical Society A,* **172**, 2, 361-381.

Durrant, G.B., Groves, R., Staetsky, L. and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviours on Refusal in Household Surveys. *Public Opinion Quarterly*, **74**, 1, 1-36.

Eckman, S. and O'Muircheartaigh, C. (2008). Optimal Subsampling Strategies in the General Social Survey. *In: Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Efron, B. (1981). Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and other Methods, *Biometrika*, **68**, 589-599.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, **7**, 325-337.

Elliott, M.R., Little, R.J.A. and Lewitzky, S. (2000). Subsampling Callbacks to Improve Survey Efficiency. *Journal of the American Statistical Association*, *Applications and Case Studies*, **95**, 451, 730-738.

Fay, M.P., Graubard, B.I., Freedman, L.S. and Midthune, D.N. (1998). Conditional Logistic Regression with Sandwich Estimators: Application to Meta Analysis. *Biometrics*, **54**, 195-208.

Foster, K. (1998). Evaluating Nonresponse on Household Surveys. *GSS Methodology Series*, 8, Office for National Statistics, London.

Fuller, W. A. (2002). Regression Estimation for Survey Samples. *Survey Methodology*, **28**, 5-23.

Fuller, W.A. (1998). Replication Variance Estimation for Two-phase Samples. *Statistica Sinica*, **8**, 1153-1164.

Fuller, W.A. (2009). *Sampling Statistics*, Hoboken: Wiley.

Fuller, W.A. and An, A.B. (1998). Regression Adjustment for Nonresponse. *Journal of the Indian Society of Agricultural Statistics,* **51**, 33, 1-342.

Fuller, W.A., Loughlin, M.M. and Baker, H.D. (1994). Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, **20**, 75-85.

Goldstein, H. (2010). *Multilevel Statistical Models*, 4th ed., Chichester: Wiley.

Goyder, J. (1987). *The Silent Minority: Nonrespondents on Sample Surveys*. Westview Press, Boulder, CO.

Goyder, J. (1994). An Experiment with Cash Incentives on a Personal Interview Survey. *Journal of the Market Research Society*, **36**, 4, 360-366.

Greenberg, B.S. and Stokes S.L. (1990). Developing an Optimal Call Scheduling Strategy for a Telephone Survey. *Journal of Official Statistics,* **6,** 4, 421-435.

Groves R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, **70**, 5, 646-675.

Groves, R.M. and Couper, M.P. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official Statistics*, **12**, 1, 63-83.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*, New York: Wiley.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society A,* **169**, 3, 439-459.

Hansen, M.H. and Hurwitz, W.N. (1946). The Problem of Non-response in Sample Surveys. *Journal of the American Statistical Association,* **41**, 517–529.

Herzog, T. N., Scheuren, F. J. and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Hopkins, K.D. and Gullickson, A.R. (1992). Response Rates in Survey Research: A Meta-Analysis of the Effects of Monetary Gratuities. *Journal of Experimental Education,* **61**, 52-62.

Hox, J.J. and de Leeuw E. (2002). The Influence of Interviewers' Attitude and Behavior on Household Survey Nonresponse : An International Comparison. In: *Survey Nonresponse*, Groves R.M., Dillman Don A., Eltinge J.L., Little R. J. A., New York: Wiley, 103-119.

Iannacchione, V.G. (2003). Sequential Weight Adjustment for Location and Cooperation Propensity for the 1995 National Survey of Family Growth. *Journal of Official Statistics*, **19**, 31-43.

Ireland, C. T. and Kullback, S. (1968). Contingency Tables with Given Marginals. *Biometrika*, **55**, 179-188.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, **19**, 81-97.

Kalton, G. and Maligalig, D.S. (1991). A Comparison of Methods for Weighting Adjustment for Nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, **29**, 119-127.

Kennickell, A. (2003). Reordering the Darkness: Application of Effort and Unit Nonresponse in the Survey of Consumer Finances. *In Proceedings of the Survey Research Methods Section, American Statistical Association*.

Kim, J.K. & Kim, J.J. (2007). Nonresponse Weighting Adjustment Using Estimated Probability. *Canadian Journal of Statistics*, **35**, 501-514.

Kish L. (1965). S*urvey Sampling*. New York: Wiley

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, **32**, 133-142.

Kulka, R.A. and Weeks, M.F. (1988). Towards the Development of Optimal Calling Protocolls for Telephone Surveys: A Conditional Probabilities Approach. *Journal of Official Statistics*, **4**, 4, 319-358.

Laflamme, F., Maydan, M. and Miller, A. (2008). Using Paradata to Actively Manage Data Collection Survey Process. *In Proceedings of the Survey Research Methods Section, American Statistical Association*.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22.

Lillard, L.A. and Panis, C.W.A. (2003). *aML Multilevel Multiprocess Statistical Software, Version 2.0*. EconWare, Los Angeles, California.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, **54**, 139-157.

Little, R.J.A. (1988). Missing Data Adjustments in Large Surveys (with discussion). *Journal Business and Economic Statistics*, **6**, 267-301.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Ed. Hoboken: Wiley.

Little, R.J.A. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**, 2, 161-168.

Lundström, S. and Särndal, C.E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, **15**, 2, 305–327.

Lynn, P., Clarke, P., Martin, J. and Sturgis, P. (2002). The Effect of Extended Interviewer Efforts on Nonresponse Bias. *In: Survey Nonresponse*, Groves R.M., Dillman Don A., Eltinge J.L., Little R. J. A., New York: Wiley, 135-147.

Matsuo, H., Loosveldt, G. and Billiet, J. (2006). The History of the Contact Procedure and Survey Cooperation – Applying Demographic Methods to European Social Survey Contact Forms Round 2 in Belgium. Louvain-la-Neuve, Belgium. *Paper presented at the Quetelet Conference*.

Morton-Williams J. (1993). *Interviewer Approaches*, Aldershot: Dartmouth.

Nicoletti, C. and Peracchi, F. (2005). Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel. *Journal of the Royal Statistical Society A*, **168**, 4, 763-781.

O'Muircheartaigh, C. and Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Nonresponse. *Journal of the Royal Statistical Society A,* **162**, 3, 437-446.

Office for National Statistics (1998). *Labour Force Survey User Guide*, *Volume1: Background and Methodology,* London.

Olson K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, **70**, 5, 737-758.

Peytchev, A. (2011). Correction for Survey Nonresponse and Measurement Error. *Paper presented at the 10th Conference on Health Survey Research Methods, April.*

Pickery, J. and Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components on Unit Nonresponse. *Quality & Quantity*, **36**, 427-37.

Pickery, J. and Loosveldt, G. (2004). A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers. *Journal of Official Statistics*, **20**, 1, 77-89.

Pickery, J. and Loosveldt, G. and Carton, A. (2001). The Effects of Interviewers and Respondents Characteristics on Response Bahavior in Panel Surveys. *Sociological Methods and Research*, **29**, 509-23.

Politz, A.N. and Simmons, W.R. (1949). An Attempt to Get 'not-at-homes' Into the Sample Without Call-backs. *Journal of the American Statistical Association*, **44**, 9-31.

Purdon, S., Campanelli, P. and Sturgis, P. (1999). Interviewer's Calling Strategies on Face-to-Face Interview Surveys. *Journal of Official Statistics,* **15**, 2, 199-216.

Quatember, A et al (2002). Structures and Analysis of Relevant National Surveys, Chapter 7, Workpackage 2 in final report on DACSEIS project to Eurostat.

Rao, J. N. K. and Tausi, M. (2004). Estimating Function Jackknife Variance Estimators Under Stratified Multistage Sampling. *Communications in Statistics - Theory and Methods*, **33**, 9, 2087-2095.

Rasbash J., Charlton C., Browne W.J., Healy M.J.R. and Cameron B. (2009a). *MLwiN version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rasbash J., Charlton C., Jones K. and Pillinger R. (2009b). *Manual Supplement to MLwiN v2.1*. Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W.J. and Goldstein, H. (2009). *A User's Guide to MLwiN v2.10*. Centre for Multilevel Modelling, University of Bristol.

Rizzo L, Kalton G, Brick J.M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse, *Survey Methodology*, **22**, 43-53.

Rodriguez, G. and Goldman, N. (2001). Improved Estimation Procedures for Multilevel Models with Binary Response: a Case-study. *Journal of the Royal Statistical Society*, Series A, **164**, 339-355.

Rosenbaum, P.R. (1987). Model-based Direct Adjustment. *Journal American Statistical Association*, **82**, 387-94.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Särndal, C-E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Chichester, England.

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

Schnell, R. and Kreuter, F. (2001). Separating Interviewer and Sampling Point Effects. *Journal of Official Statistics*, **21**, 3, 389-410.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Shao, J. (2007). Handling Survey Nonresponse in Cluster Sampling. *Survey Methodology*, **33**, 81-85.

Singer, E. (2002). The Use of Incentives to Reduce Non Response. *In: Survey Nonresponse*, Groves R.M., Dillman Don A., Eltinge J.L., Little R. J. A., New York: Wiley, 163-177.

Singer, E. and Kulka, R.A. (2002). Paying Respondents for Survey Participation. *In Studies of Welfare Populations: Data Collection and Research Issues*, Ver Ploeg M., Moffitt R.A. and Citro C.F., National Academy Press, Washington, D.C., 105-127.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T. and McGonagle, K. (1999). The Effects of Incentives on Response Rates in Interviewer-Mediated Surveys. *Journal of Official Statistics*, **15**, 2, 217-230.

Skinner, C. and D'Arrigo, J. (2011). Inverse Probability Weighting for Clustered Nonresponse. *Biometrika*, **98**, 4, 953-966.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis*. Sage

Spiegelhalter, D. J., Best, N.G., Carlin, B. P. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society B*, **64**, 583-639.

Steele, F., Goldstein, H. and Browne, W. (2004). A General Multistate Competing Risks Model for Event History Data, with an Application to a Study of Contraceptive Use Dynamics. *Statistical Modelling*, **4**, 2, 145-159.

Stoop, I.A.L. (2005). *The hunt for the Last Respondent: Nonresponse in Sample Surveys*. Social and Cultural Planning Office, The Hague.

Stukel, D. M., Hidiroglou M. A. and Särndal, C-E. (1996). Variance Estimation for Calibration Estimators: a Comparison of Jackknifing versus Taylor Linearization. *Survey Methodology*, **22**, 117-25.

Swires-Hennessy, E. and Drake, M. (1992). The Optimum Time at Which to conduct Interviews. *Journal of the Market Research Society*, **34**, 1, 61-72.

Tan, Z. (2010). Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting. *Biometrika*, **97**, 661–682.

Tourangeau, R. (2004). Survey Research and Societal Change. *Annual Review of Psychology*, **55**, 775-801.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*, New York: Springer.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, **88**, 89-96.

Valliant, R. (2004). The Effect of Multiple Weighting Steps on Variance Estimation. *Journal of Official Statistics,* **20**, 1-18.

von Sanden, N. D. (2004). *Interviewer Effects in Household Surveys: Estimation and Design*, PhD Thesis, University of Wollongong.

Weber, D. and Burt, R. (1972). *Who's Home When*. Washington, D.C.: U.S. Bureau of the Census.

Weeks, M.F., Jones, B.L., Folsom, R.E. and Benrud, C.H. (1980). Optimal Times to Contact Sample Households. *Public Opinion Quarterly*, **44**, 1, 101-114.

Weeks, M.F., Kulka, R.A. and Pierson, S.A. (1987). Optimal Call Scheduling for a Telephone Survey. *Public Opinion Quarterly,* **51**, 4, 540-549.

White, A., Freeth, S. and Martin, J. (2001). Evaluation of Survey Data Quality Using Matched Census-Survey Records. *International Conference Quality in Official Statistics, Stockholm*, May.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd ed. Springer, New York.

Wood, A.M. and White, I.R. (2006). Using Number of Failed Contact Attempts to Adjust for Non-ignorable Non-response. *Journal of the Royal Statistical Society A,* **169**, 3, 525-542.

Yuan, Y. and Little, R.J.A. (2007). Model-based Estimates of the Finite Population Mean for Two-stage Cluster Samples with Unit Non-response. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56, 79-97.

Yuan, Y. and Little, R.J.A. (2008). Model-based Inference for Two-stage Cluster Samples Subject to Nonignorable Item Nonresponse, *Journal of Official Statistics*, **24**, 193-211.

Yung, W., Rao, J.N.K. (2000). Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information. *Journal of the American Statistical Association*, **95**, 903-915.