Joint Calibration Estimator for Dual Frame Surveys

by

Mahmoud Ahmed Elkasabi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2013

Doctoral Committee:

Research Scientist Steven G. Heeringa, Co-Chair
Research Professor James M. Lepkowski, Co-Chair
Assistant Research Scientist Sunghee Lee
Research Professor Richard L. Valliant

Dedications

This dissertation is dedicated to Hala and Salma.

Hala, thanks for your continuing support, patience and encouragement throughout this work and throughout our journey together.

Salma, you bring joy to my life. Your birth was a great motivation and bright inspiration to finish this work.

Acknowledgements

I would like to extend my utmost gratitude to my co-chairs, Steve Heeringa and Jim Lepkowski, who believed in my dissertation ideas and gave me from their time. Steve and Jim have provided valuable advice on every aspect of my professional life. It has been a pleasure to work with both of them. I am also grateful for my committee members, Richard Valliant and Sunghee Lee, who gave me extensive feedback on my dissertation ideas, at an early stage of developing them. Also I would like to thank my friends, colleagues and faculties for their support and help over the years in the Program in Survey Methodology (PSM) and the Joint Program in Survey Methodology (JPSM).

Table of Contents

List of Figures

List of Tables

Abstract

Dual frame surveys are becoming more common in survey practice due to rapid changes in the cost of survey data collection, as well as changes in population coverage patterns and sample unit accessibility. Many dual frame estimators have been proposed in the literature. Some of these estimators are theoretically optimal but hard to be applied in practice, whereas the rest are applicable but not as optimal as the first group. All the standard dual frame estimators require accurate information about the design domain membership.

In this dissertation, a set of desirable properties for the dual frame estimators is specified. These properties are used as criteria to evaluate the standard dual frame estimators. At the same time, the Joint Calibration Estimator (JCE) is proposed as a new dual frame estimator that is simple to apply and meets most of the desirable properties for dual frame estimators.

In Chapter 2, the JCE is introduced as an approximately unbiased dual frame estimator, with a degree of unbiasedness depending on the relationship between study variables and auxiliary variables. The JCE achieves better performance when the auxiliary variables can fully explain the variability in the study variables of interest or at least when the auxiliary variables are strong correlates of the study variables. The JCE for point estimates can be applied by standard survey software and can easily be extended to multiple frame survey estimation. In Chapter 3, the JCE properties are explored in the presence of the nonresponse error. Theoretically and empirically, the JCE proves to be robust to nonresponse error as long as a strong set of auxiliary

variables is used. This strong set should predict both the response mechanism and the main study variables.

Finally, the effect of domain misclassification on the dual frame estimators is discussed in Chapter 4. Since the JCE does not require domain membership information, it tends to be robust against domain misclassification even if domain totals are included in the calibration auxiliary variables.

Chapter 1

Dual Frame Samples Estimation

1.1 Introduction

Historically, dual frame designs have been used to achieve better population coverage at lower survey cost than single frame alternatives. The early applications of the dual frame designs, or multiple frame designs in general, were in business surveys such as the Sample Survey of Retail Stores (Hansen, Hurwitz, & Madow, 1953) and in agriculture surveys (González-Villalobos & Wallace, 1996). Dual frame area-landline surveys, composed of an area frame and Random-Digit-Dialing (RDD) landline telephone frame, were shown to achieve better population coverage at lower cost (Lepkowski & Groves, 1986). With rapid changes in the cost of survey data collection, changes in population coverage patterns and sample unit accessibility, dual frame sample surveys are becoming more common in survey practice. For example, dual frame telephone surveys that combine RDD landline telephone samples and cell phone samples emerged to reduce the noncoverage of “cell-only” households in RDD landline telephone surveys (Brick et al., 2007; Blumberg & Luke, 2011; Keeter, 2006; Keeter, Kennedy, Clark, Tompson, & Mokrzycki, 2007; Link, Battaglia, Frankel, Osborn, & Mokdad, 2007). At the same time, Address-Based-Sampling (ABS) sampling has been explored as a complement or an alternative to the RDD telephone surveys in a number of recent studies (Link, Battaglia, Frankel, Osborn, & Mokdad, 2006, 2008; Link & Lai, 2011).

In dual frame surveys, the intersection between the two frames can be non-overlapping (Figure 1.1), partially overlapping (Figure 1.2) or completely overlapping (Figure 1.3) (Lohr, 2009, 2011). In non-overlapping dual frame designs, the estimation is straightforward since the sampling plan can be considered as a stratified sample with two strata. However, in the overlapping designs, the estimation is not as straightforward. Due to the overlap, simply adding the two samples' estimated totals results in a biased estimate of the overall total. Standard dual frame estimators adjust for the overlap or multiplicity in the intersecting domain (Lohr, 2011).

The standard dual frame estimators present many methodological and practical problems in their implementation (Lohr, 2011). These problems can delay the processing of "quick turn-around" surveys. At the same time, the correct identification of the design domain for each sample element is essential. Therefore, non-sampling errors in the determination of design domain membership can affect the efficiency of estimates (Lohr, 2011; Mecatti, 2007).

Figure 1.1: Non-overlapping frames A and B

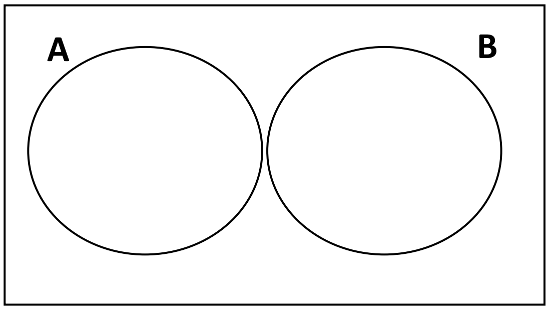


Figure 1.2: Partially overlapping frames A and B with three domains *a*, *b* and *ab*

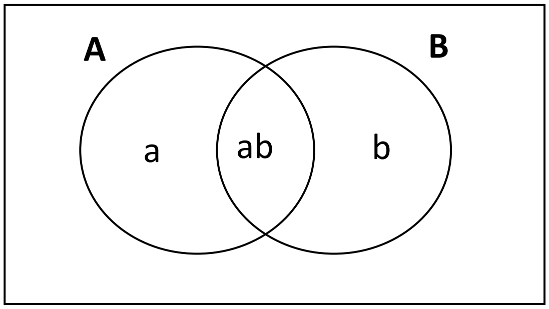
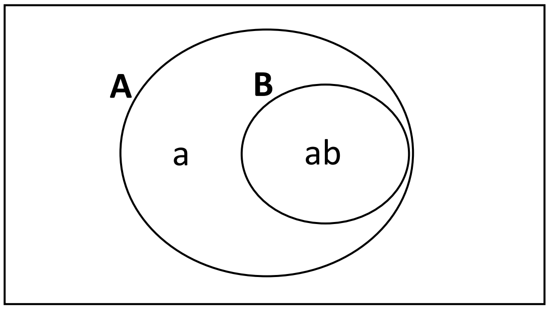


Figure 1.3: Completely overlapping frames A and B with two domains *a* and *ab*



Since dual frame designs have become more common in practice, it has been important to find simple, yet efficient dual frame estimators that can be applied easily in survey practice, with fewer requirements and comparable or better efficiency, compared to standard dual frame estimators. This chapter provides background information on the standard dual frame estimators and their properties. Desirable properties of estimators from dual frame surveys are discussed in Section 1.2. An overview of the standard dual frame estimators in the context of the desirable properties is discussed in Sections 1.3 and 1.4. Conclusions and motivations for the dissertation are presented and discussed in Section 1.5.

1.2 Desirable Properties for Dual Frame Estimators

Lohr (2011) identified the following five desirable properties for dual frame estimators. (1) An estimator should be unbiased for the corresponding finite population quantity. (2) An estimator should be internally consistent; that is, the multivariate relationships in the data should be preserved. For example, the sum of the estimated totals for male and female smokers should be equal to the estimated total for all smokers. Internal inconsistency may happen if estimators are dependent on study variable *y*, requiring a different set of weights for each study variable. (3) An estimator should be efficient, with low Mean Square Error (MSE). (4) An estimator should be of a form that can be calculated with standard survey software. This means that only one set of weights is available for all study variables and design variable or replicate weights are available for formula-based or replication-based variance estimation, respectively. (5) An estimator should be robust to non-sampling errors.

In addition to the previous properties, the following three properties should be added. (6) Data requirements for estimator should be reasonable. For example, information about design domain membership or variance and covariance components might be required for some estimators, but only poorly measured or unreliable components could be available in practice, which adds to the burden of computing the estimator. (7) An estimator should be robust to non-sampling errors in the estimator's requirements. For example, although some estimators might theoretically be efficient, reporting errors in the required information about design domain membership or biased estimates of the required variance and covariance components could result in biased or non-optimal estimators. And finally, (8) an estimator should be applicable for dual frame and multiple frame surveys. Since most of the multiple frame estimators are proposed for dual frame surveys, the previous properties should be explored in the context of multiple frame surveys, as well. As will be discussed in the next chapters, the last three properties are the primary motivation for the current study.

1.3 Dual Frame Estimators for Surveys

Let $U = \{1,..,k,..,N\}$ denote a target population of N elements, and let $U_A = \{1,..,k,..,N_A\}$ and $U_B = \{1,..,k,..,N_B\}$ denote two overlapping sub-populations. The two sub-populations are not assumed to be exclusive, that is: $U_A \bigcap U_B = U_{ab} \neq 0$ and $U_A \bigcup U_B = U$. The dual frame design sample s is composed of two samples $s_A \left(s_A \subseteq U_A\right)$ and $s_B \left(s_B \subseteq U_B\right)$ selected from the two overlapping populations U_A and U_B using a sample design with inclusion probabilities $\pi_k^A = p\left(k \in s_A\right)$ and $\pi_k^B = p\left(k \in s_B\right)$, where the frame populations agree with the target populations. Base weights to compensate for unequal selection probabilities are denoted by $d_k = \left(d_k^A, d_k^B\right)$, $d_k^A = 1/\pi_k^A$ for s_A and by $d_k^B = 1/\pi_k^B$ for s_B. Let N_A and N_B denote the population sizes and n_A and n_B denote the sample sizes for frames A and B, respectively. Let $a = A \cap B^c$, $b = A^c \cap B$ and $ab = A \cap B$, where c denotes complement of a set, and $s_a = a \cap s_A$, $s_b = b \cap s_B$, $s_{ab}^A = ab \cap s_A$ and $s_{ab}^B = ab \cap s_B$. Most of the standard dual frame estimators of a population total take the form

$$\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b \qquad (1.1)$$

to estimate the true population total $Y = Y_a + Y_{ab} + Y_b$.

A standard estimation method can be used to find domain a and b estimates of totals $\hat{Y}_a = \sum_{s_a} d_k y_k$ and $\hat{Y}_b = \sum_{s_b} d_k y_k$ for a population characteristic, Y. However, to find $\hat{Y}_{ab}$, consider the estimators $\hat{Y}_{ab}^A = \sum_{s_{ab}^A} d_k y_k$ and $\hat{Y}_{ab}^B = \sum_{s_{ab}^B} d_k y_k$. For each sample, the estimators of population totals are unbiased for the corresponding domain total Y_a, Y_{ab} and Y_b, $E\left[\hat{Y}_a + \hat{Y}_{ab}^A\right] \approx Y_a + Y_{ab}$ and $E\left[\hat{Y}_b + \hat{Y}_{ab}^B\right] \approx Y_b + Y_{ab}$. Therefore adding the two sample estimated totals results in a biased population estimate

$$E\left[\hat{Y}_a + \hat{Y}_{ab}^A + \hat{Y}_b + \hat{Y}_{ab}^B\right] \approx Y_a + 2Y_{ab} + Y_b \neq Y \tag{1.2}$$

Finding an unbiased dual frame estimator for *Y* can be accomplished by using a weighted average of the estimators $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$. The unbiased dual frame estimator can take the form

$$\hat{Y} = \hat{Y}_a + \theta\hat{Y}_{ab}^A + \left(1-\theta\right)\hat{Y}_{ab}^B + \hat{Y}_b \tag{1.3}$$

where $\theta \in \left[0,1\right]$ is a composite factor combining $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$. Estimators of domain sizes $\hat{N}_a^A$, $\hat{N}_{ab}^A$, $\hat{N}_{ab}^B$ and $\hat{N}_b^B$ are defined by setting $y_k = 1$ for all $k = 1,...,n$ in $\hat{Y}_a^A$, $\hat{Y}_{ab}^A$, $\hat{Y}_{ab}^B$ and $\hat{Y}_b^B$, and the dual frame estimator in (1.3) can be used to find the population total estimate $\hat{N}$. Consequently, an unbiased dual frame estimator for population mean $\overline{Y}$ can be written as $\overline{Y} = \hat{Y}/\hat{N}$. The weighted version of the estimated total in (1.3) can be written as

$$\hat{Y} = \sum\nolimits_{s_A} m_k d_k y_k + \sum\nolimits_{s_B} m_k d_k y_k = \sum\nolimits_{s_A} w_k y_k + \sum\nolimits_{s_B} w_k y_k \tag{1.4}$$

where w_k is a final weight. The adjustment factor m_k can be written as

$$m_k = \begin{cases} 1 & k \in s_a \\ \theta & k \in s_{ab}^A \\ 1-\theta & k \in s_{ab}^B \\ 1 & k \in s_b \end{cases} \tag{1.5}$$

Until recently, the weighted version in (1.4) has not been explicitly defined in the literature. However, in a comprehensive chapter on multiple frame surveys, Lohr (2009) has set out the weighted version for the different estimators. The explicit weighted version is essential for (1) application of standard survey software, and (2) finding estimators other than totals and means, such as the ratio estimator $\hat{R} = \sum_s w_k y_k \big/ \sum_s w_k x_k$ or the simple linear regression coefficient estimator $\hat{\beta} = \sum_s w_k x_k y_k \big/ \sum_s w_k x_k^2$.

It is worth noting that all dual frame estimators in the following sub-sections are approximately unbiased, the first desirable property for multiple frame survey estimators. At the same time, with regard to the fifth property, the effect of non-sampling errors may be qualitatively different from those in single frame surveys (Brick, Flores-Cervantes, Lee, & Norman, 2011) because non-sampling errors may causally associate with the sampling frame. Also, sampling from more than one frame results in non-sampling errors with differential effects, adding to the complexity of the assessment and adjustment for these errors. Finally, most of the following estimators require accurate information about domain membership, which might be affected by reporting errors and, in turn, leading to a biased $\hat{Y}$.

1.3.1 Hartley Estimator (HE)

The standard Hartley (1962) dual frame estimator (HE) for the estimated total of y can be written as

$$\hat{Y}_{HE} = \hat{Y}_a + \theta_{HE}\hat{Y}_{ab}^{A} + \left(1-\theta_{HE}\right)\hat{Y}_{ab}^{B} + \hat{Y}_b \qquad (1.6)$$

Hartley (1962, 1974) proposed choosing the composite factor θ_{HE} to minimize the variance of $\hat{Y}_{HE}$. The optimizing value of θ_{HE} can be written as

$$\theta_{HE} = \frac{V\left(\hat{Y}_{ab}^{B}\right) + Cov\left(\hat{Y}_{b}^{B}, \hat{Y}_{ab}^{B}\right) - Cov\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A}\right)}{V\left(\hat{Y}_{ab}^{A}\right) + V\left(\hat{Y}_{ab}^{B}\right)} \tag{1.7}$$

Generally, the components of (1.7) are unknown and need to be estimated from the data. An estimated version of θ_{HE} can be written as

$$\hat{\theta}_{HE} = \frac{\hat{V}\left(\hat{Y}_{ab}^{B}\right) + \widehat{Cov}\left(\hat{Y}_{b}^{B}, \hat{Y}_{ab}^{B}\right) - \widehat{Cov}\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A}\right)}{\hat{V}\left(\hat{Y}_{ab}^{A}\right) + \hat{V}\left(\hat{Y}_{ab}^{B}\right)} \tag{1.8}$$

The weighted version of $\hat{Y}_{HE}$ can be written as in (1.4) with the modification factor

$$m_k = \begin{cases} 1 & k \in s_a \\ \hat{\theta}_{HE} & k \in s_{ab}^{A} \\ 1 - \hat{\theta}_{HE} & k \in s_{ab}^{B} \\ 1 & k \in s_b \end{cases} \tag{1.9}$$

The domain post-stratified version of $\hat{Y}_{HE}$ can be written as

$$\hat{Y}_{HE}^{post} = \frac{N_a}{\hat{N}_a}\hat{Y}_a + \frac{N_{ab}}{\hat{N}_{ab}}\left(\hat{\theta}_{HE}\hat{Y}_{ab}^{A} + \left(1 - \hat{\theta}_{HE}\right)\hat{Y}_{ab}^{B}\right) + \frac{N_b}{\hat{N}_b}\hat{Y}_b \tag{1.10}$$

where N_a , N_{ab} and N_b denote the population sizes for domains *a*, *ab* and *b*, respectively. $\hat{N}_a = \sum_{s_a} d_k$ and $\hat{N}_b = \sum_{s_b} d_k$ are the estimated non-overlapping domain totals and

$\hat{N}_{ab} = \hat{\theta}_{HE}\hat{N}_{ab}^{A} + \left(1-\hat{\theta}_{HE}\right)\hat{N}_{ab}^{B}$ is the overlapping domain estimated total, where $\hat{N}_{ab}^{A} = \sum_{s_{ab}^{A}} d_k$ and $\hat{N}_{ab}^{B} = \sum_{s_{ab}^{B}} d_k$.

Although $\hat{Y}_{HE}$ can be classified as an efficient estimator, it is internally inconsistent since it generates weights that are dependent on the study variables, y. This restricts the practical application of HE using standard survey software. HE also requires accurate estimates of variance and covariance components for finding composite factor, $\hat{\theta}_{HE}$. Biased estimates of the required variance and covariance components could result in non-optimal $\hat{Y}_{HE}$. Finally, deriving HE for multiple frame surveys is also more complicated due to the need to estimate the covariance terms required for composite factors, θ_{HE}.

1.3.2 Fixed Weight Estimator (FWE)

Choosing an arbitrary fixed value $\theta \in [0,1]$ for the composite factor in (1.3) (e.g. θ=0.5) yields the unbiased Fixed Weight Estimator (FWE), which depending on the arbitrary choice may or may not be as efficient as the HE (Hartley, 1962). The post-stratified version of $\hat{Y}_{FWE}$ can be written as in (1.10) with the fixed value composite factor $\theta \in [0,1]$ instead of θ_{HE}. The FWE is internally consistent and results in only one set of weights for all study variables. Also, deriving FWE for multiple frame surveys is straightforward. The weighted version of FWE can be written as in (1.4) and (1.5), with the fixed value composite factor $\theta \in [0,1]$.

1.3.3 Fuller-Burmeister Estimator (FB)

Fuller and Burmeister (1972) extended Hartley's dual-frame estimator of population totals by considering information about the maximum likelihood estimator $\hat{N}_{ab}$ of the overlap domain population size N_{ab}. The Fuller-Burmeister estimator (FB) of total of y can be written in the form of estimated domain totals as

$$\hat{Y}_{FB} = \hat{Y}_a^A + \theta_{1\mathrm{FB}} \hat{Y}_{ab}^A + \left(1 - \theta_{1\mathrm{FB}}\right) \hat{Y}_{ab}^B + \hat{Y}_b^B + \theta_{2\mathrm{FB}} \left(\hat{N}_{ab}^A - \hat{N}_{ab}^B \right) \tag{1.11}$$

The optimal values of $\theta_{1\mathrm{FB}}$ and $\theta_{2\mathrm{FB}}$ are chosen to minimize the variance of $\hat{Y}_{FB}$. Skinner (1991) desctibed $\hat{Y}_{FB}$ as a maximum likelihood estimator (MLE) since it can be derived from maximum likelihood (ML) principles. Although it is an efficient estimator with a small asymptotic variance, the FB estimator is, like the HE, study-variable dependent and internally inconsistent. Therefore, the FB cannot be applied using the standard survey software. Finding the optimal values of $\theta_{1\mathrm{FB}}$ and $\theta_{2\mathrm{FB}}$ requires estimation of the variance and covariance components. Biased estimates of these components could result in a non-optimal estimator $\hat{Y}_{FB}$. Deriving FB for multiple frame surveys is also more complicated due to the need to estimate the covariance terms required for composite factors, $\theta_{1\mathrm{FB}}$ and $\theta_{2\mathrm{FB}}$. Finally, the HE can be considered a special case of the FB where the composite factor $\theta_{2\mathrm{FB}} = 0$.

1.3.4 Single Frame Estimator (SFE)

Bankier (1986) and Kalton and Anderson (1986) proposed the Single Frame Estimator (SFE) which treats the dual frame design as a single frame design. The SFE treats the dual frame design as a stratified design composed of three strata, one for each design domain, estimating joint inclusion probabilities. Provided domain membership is known, SFE weights are much easier to calculate than the HE and the FB estimators (Bankier, 1986). Under an assumption that the probability of duplicate sample selection from the separate frames is negligible, the single frame estimator $\hat{Y}_{SFE}$ can be written as the weighted version in (1.4), with adjustment factor

$$m_k = \begin{cases} 1 & k \in s_a \\ d_k^B\left(d_k^A + d_k^B\right)^{-1} & k \in s_{ab}^A \\ d_k^A\left(d_k^A + d_k^B\right)^{-1} & k \in s_{ab}^B \\ 1 & k \in s_b \end{cases} \tag{1.12}$$

Under simple random sampling, or other self-weighting sample design, the SFE is a special case of the HE where the composite factor $\theta_{HE} = d_k^B\left(d_k^A + d_k^B\right)^{-1}$. Raking ratio or regression estimation can be used to adjust the SFE (Bankier, 1986; Lohr & Rao, 2000; Rao & Skinner, 1996; Skinner, 1991). Under simple self-weighting sample designs, adjusting the inclusion probabilities in the overlapping domains by SFE adjustment factors is straightforward. However, this adjustment is complicated under complex sampling designs, such as stratified samples, because adjusting the inclusion probabilities for an overlapping domain case selected in frame A requires knowing the inclusion probability of the same case in frame B.

$\hat{Y}_{SFE}$ is not an efficient estimator, but it is internally consistent since it generates only one set of weights for all study variables, y. The standard survey software can be used to find $\hat{Y}_{SFE}$. Deriving SFE for multiple frame surveys, under simple random sampling plans, is straightforward. It is, however, more complicated with complex sampling plans.

1.3.5 Pseudo-Maximum Likelihood Estimator (PML)

Skinner and Rao (1996) extended the maximum likelihood FB estimator to achieve internal and design-based consistency under complex designs and developed the Pseudo-Maximum Likelihood Estimator (PML). Under unknown N_{ab}, the pseudo-maximum likelihood estimator, $\hat{N}_{ab}^{PML}$, can be derived as the smaller of the roots of the quadratic equation

$$\left[\frac{\theta_P}{N_B}+\frac{1-\theta_P}{N_A}\right]x^2-\left[1+\theta_P\frac{\hat{N}_{ab}^{A}}{N_B}+\left(1-\theta_P\right)\frac{\hat{N}_{ab}^{B}}{N_A}\right]x+\theta_P\hat{N}_{ab}^{A}+\left(1-\theta_P\right)\hat{N}_{ab}^{B}=0 \qquad (1.13)$$

The PML estimator can be written as

$$\hat{Y}_{PML}=\frac{N_A-\hat{N}_{ab}^{PML}}{\hat{N}_a}\hat{Y}_a^A+\frac{N_B-\hat{N}_{ab}^{PML}}{\hat{N}_b}\hat{Y}_b^B+\frac{\hat{N}_{ab}^{PML}}{\theta_P\hat{N}_{ab}^{A}+\left(1-\theta_P\right)\hat{N}_{ab}^{B}}\left(\theta_P\hat{Y}_{ab}^{A}+\left(1-\theta_P\right)\hat{Y}_{ab}^{B}\right) \quad (1.14)$$

where θ_P is chosen to minimize the variance of $\hat{N}_{ab}^{PML}$ as

$$\theta_P=\frac{N_aN_BV\left(\hat{N}_{ab}^{B}\right)}{N_aN_BV\left(\hat{N}_{ab}^{B}\right)+N_bN_AV\left(\hat{N}_{ab}^{A}\right)} \qquad (1.15)$$

The weighted version of PML $\hat{Y}_{PML}$ can be written as in (1.4) with the adjustment factor

$$m_k=\begin{cases}\dfrac{N_A-\hat{N}_{ab}^{PML}}{\hat{N}_a^{A}} & k\in s_a\\[2ex] \dfrac{\hat{N}_{ab}^{PML}}{\hat{\theta}_p\hat{N}_{ab}^{A}+\left(1-\hat{\theta}_p\right)\hat{N}_{ab}^{B}}\hat{\theta}_p & k\in s_{ab}^{A}\\[2ex] \dfrac{\hat{N}_{ab}^{PML}}{\hat{\theta}_p\hat{N}_{ab}^{A}+\left(1-\hat{\theta}_p\right)\hat{N}_{ab}^{B}}\left(1-\hat{\theta}_p\right) & k\in s_{ab}^{B}\\[2ex] \dfrac{N_B-\hat{N}_{ab}^{PML}}{\hat{N}_b^{B}} & k\in s_b\end{cases} \qquad (1.16)$$

$\hat{Y}_{PML}$ is both an internally consistent and an efficient estimator. Standard survey software can be used to compute $\hat{Y}_{PML}$. However, PML requires accurate estimates of variance components for calculating the composite factor, θ_{p}. Biased estimates of the required variance components could result in non-optimal $\hat{Y}_{PML}$. Deriving PML for multiple frame surveys is also more complicated due to the need to estimate the covariance terms required for composite factors, θ_{p}.

1.3.6 Multiplicity Estimator (ME)

Since some population elements have multiple opportunities to be selected as sample elements, estimation of population statistics from multiple frame surveys can in general be formalized as a *multiplicity* problem. Meccati (2007) introduced a simple dual frame estimator, the Multiplicity Estimator (ME), which depends on the multiplicity information, M_k, the number of frames that case *k* belongs to. The multiplicity estimator for multiple frames design with *Q* overlapping frames can be written as

$$\hat{Y}_{ME} = \sum_{Q} \sum_{s_q} M_k^{-1} w_k^q y_k \tag{1.17}$$

The weighted version of ME can be written as in (1.4) and (1.5) with composite factor $\theta = M_k^{-1}$. This estimator comes under the Generalized Multiplicity-adjusted Horvitz-Thompson (GMHT) approach proposed by Singh & Mecatti (2011). Beside the flexibility of extending the ME to general multiple frame designs, only partial multiplicity information is required for this estimator. It can be obtained by asking the sampled unit *how many frames they belong to*. Under dual frame designs, both the ME and the FWE are equivalent where the composite factor $\theta = 0.5$. Both can be considered as special cases of the HE.

$\hat{Y}_{ME}$ is an internally consistent, which can be computed by standard survey software, but inefficient estimator. ME does not require full information about specific design domain membership; only partial information about total memberships M_k is enough. Thus, deriving ME for multiple frame surveys is a straightforward.

1.3.7 Pseudo-empirical Likelihood Estimator (PEL)

Rao and Wu (2010) proposed the Pseudo-empirical likelihood (PEL) estimator, which depends on finding the adjustment factor m_k based on maximizing an empirical log likelihood function which can be written as

$$\begin{aligned}\ell\left(\mathrm{p}_a, \mathrm{p}_{ab}^A, \mathrm{p}_{ab}^B, \mathrm{p}_b\right) = \frac{n_A + n_B}{N}\Bigg[&\sum\nolimits_{s_a} \frac{N_a}{\hat{N}_a} w_k \log\left(p_{ak}\right) + \sum\nolimits_{s_{ab}^A} \frac{\theta_p N_{ab}}{\hat{N}_{ab}^A} w_k \log\left(p_{abk}^A\right) \\ &+ \sum\nolimits_{s_b} \frac{N_b}{\hat{N}_b} w_k \log\left(p_{bk}\right) + \sum\nolimits_{s_{ab}^B} \frac{\left(1-\theta_p\right) N_{ab}}{\hat{N}_{ab}^B} w_k \log\left(p_{abk}^B\right)\Bigg]\end{aligned} \tag{1.18}$$

where $\mathrm{p}_a = \left(p_{a1},....,p_{an_a}\right)'$, $\mathrm{p}_{ab}^A = \left(p_{ab1}^A,....,p_{abn_{ab}^A}^A\right)'$, $\mathrm{p}_b = \left(p_{b1},....,p_{bn_b}\right)'$ and

$\mathrm{p}_{ab}^B = \left(p_{ab1}^B,....,p_{abn_{ab}^B}^B\right)'$ are probability measures corresponding to poststratified samples s_a, s_{ab}^A, s_b and s_{ab}^B, respectively, and θ_P can be obtained as

$$\theta_P = \frac{N_a N_B V\left(\hat{N}_{ab}^B\right)}{N_a N_B V\left(\hat{N}_{ab}^B\right) + N_b N_A V\left(\hat{N}_{ab}^A\right)} \tag{1.19}$$

subject to the constraints

$$\begin{aligned}
&\sum\nolimits_{s_a} p_{ak} = 1,\ \ \sum\nolimits_{s_{ab}^{A}} p_{abk}^{A} = 1\\
&\sum\nolimits_{s_b} p_{bk} = 1,\ \ \sum\nolimits_{s_{ab}^{B}} p_{abk}^{B} = 1\\
&\text{and}\\
&\sum\nolimits_{s_{ab}^{A}} p_{abk}^{A} y_k = \sum\nolimits_{s_{ab}^{B}} p_{abk}^{B} y_k
\end{aligned} \tag{1.20}$$

The weighted version of PEL can be written as in (1.4) with the modification factor

$$m_k = \begin{cases} \dfrac{p_{ai}}{w_i}\left(N_A - \hat{N}_{ab}^{PML}\right) & k \in s_a \\ \hat{\theta}_p \dfrac{p_{abi}^{A}}{w_i} \hat{N}_{ab}^{PML} & k \in s_{ab}^{A} \\ \left(1-\hat{\theta}_p\right) \dfrac{p_{abi}^{B}}{w_i} \hat{N}_{ab}^{PML} & k \in s_{ab}^{B} \\ \dfrac{p_{bi}}{w_i}\left(N_B - \hat{N}_{ab}^{PML}\right) & k \in s_b \end{cases} \tag{1.21}$$

PEL can be classified as an efficient estimator, but, as in the case of the HE and the FB, it is study-variable dependent and internally inconsistent. This restricts the practical application of PEL using standard survey software. As in the case of PML, PEL requires accurate estimates of variance components for finding the composite factor, θ_{p}. Biased estimates of the required variance components could result in non-optimal $\hat{Y}_{PEL}$. Deriving PEL for multiple frame surveys is also more complicated due to the need to estimate the covariance terms required for composite factors, θ_{p}.

1.4 Variance Estimation

Except for PML, variance estimation is straightforward for the internally consistent dual frame estimators (FWE, SFE and ME) that produce only one set of weights. In this case the weight adjustment factor does not depend on the individual study variable, y. Therefore, the dual frame variance can be estimated by adding the estimated variances of the estimators for two samples as

$$V\left[\hat{Y}\right]=V\left[\sum_{s_A} m_k d_k y_k\right]+V\left[\sum_{s_B} m_k d_k y_k\right] \qquad (1.22)$$

However, for PML and the internally inconsistent dual frame estimators HE, FB and PEL, the variability in the estimated composite factor θ must be captured in the variance estimation. The variability added by calibrating the design weights to population totals should also be considered in the variance estimation. For these estimators, PML, internally inconsistent estimators and calibrated dual frame estimators in general, jackknife and bootstrap methods are recommended for variance estimation (Lohr, 2011; Lohr & Rao, 2000; Skinner & Rao, 1996).

1.5 Conclusions and Motivations

1.5.1 Conclusions

The standard dual frame estimators can be classified into three groups. The first group includes the optimal estimators HE, FB and PEL, which have optimal theoretical properties but present methodological and practical problems due to their complexity, especially in the case of the multiple frame surveys (Lohr & Rao, 2000, 2006; Mecatti, 2007; Skinner, 1991). The second group includes the estimators FWE, SFE and ME which are readily calculated in practice but achieve lower efficiency relative to the optimal estimators. The third group has a single estimator, the pseudo-optimum estimator PML. It is a balance between the previous groups. PML has more practical applicability than the optimal estimators in the first group, and better efficiency than the practical estimators in the second group (Lohr, 2011; Lohr & Rao, 2000; Skinner & Rao, 1996). PML has a smaller mean square error (MSE) than FB and HE, since the variability in estimating the components of the composite factor in FB and HE adds to the estimated variance in MSE (Skinner & Rao, 1996).

Most of the standard dual frame estimators require accurate information about domain membership (multiplicity information). If this information is not available before the data collection (e.g. from properties or actual matching of frames), multiplicity information for each sample unit should be collected during the interview. Ideally, such information on frame membership should be free from reporting bias or measurement errors, but this is not typically the case in practice (Lohr & Rao, 2006). Moreover, the rate of nonresponse or misreporting for the domain membership questions could be even higher when surveying sensitive characteristics or elusive populations (Mecatti, 2007). Such problems in measuring the domain membership can have a direct effect on the error properties of the dual frame estimator.

1.5.2 Motivation

The previous discussion indicates that there is still a need for a dual frame estimator that satisfies more the desirable properties discussed earlier. In this dissertation, a new dual frame estimator will be introduced and evaluated in the context of these desirable properties. This estimator depends on the general calibration approach introduced by Deville and Sarndal (1992).

In the literature, dual frame samples can be calibrated separately, before combining the two samples, or jointly, after combining the two samples. However, the implicit potential of jointly calibrating dual frame samples has not been explored.

Since calibration generates unbiased auxiliary variable estimates under dual frame designs, there is an interest in developing and testing the calibration effect on the study variables estimates. In the following chapters, the Joint Calibration Estimator (JCE) will be introduced as opposed to the standard dual frame estimators discussed in this chapter. The JCE will be empirically compared to the standard FWE estimator. Comparisons will be made under a full response assumption and in the presence of non-sampling errors arising from survey nonresponse or domain misclassification.

Chapter 2

Joint Calibration Estimator for Dual Frame Surveys

2.1 Introduction

Chapter 1 presented a review of past and current literature on dual frame estimation. In this chapter, we provide an overview of calibration weighted estimation and introduce the model-assisted design-based Joint Calibration Estimator (JCE) for dual frame estimation. The properties of the JCE are explored under the 'ideal situation' in which non-sampling errors are absent. The calibration approach is discussed in Section 2.2. The JCE is introduced in Sections 2.3 and 2.4. In Sections 2.5 and 2.6 a bias and a variance estimator for JCE are derived. The performance of JCE in comparison with one of the dual frame estimators presented in Chapter 1 is explored in a simulation study described in Section 2.7. The simulation results and findings are presented and discussed in Sections 2.8 and 2.9.

2.2 The Calibration Approach

A standard weighting procedure in both single and dual frame designs is to rake or post-stratify weights to external population control totals. In dual frame surveys, raking or post-stratification can be performed before combining the two samples to adjust for, say, differential nonresponse in the samples from the separate frames (Brick et al., 2011). Also, these techniques can be performed after combining the two samples to retrieve properties of the original sampling weights lost in the combining step (Lohr, 2011; Lohr & Rao, 2000).

Raking and post-stratification are special cases of *calibration* adjustment. Calibration can be conceptualized as a method of constraining the weights by conditioning on auxiliary variable distributions (Deville & Särndal, 1992; Deville, Särndal & Sautory, 1993). A comprehensive description of calibration weighting methods can be found in (Särndal, 2007).

In the single frame survey design, where the sample $s\left(s\subseteq U\right)$ is selected from the population U using a sample design with inclusion probability of $\pi_k = p\left(k\in s\right)$, the base weights are denoted by $d_k = 1/\pi_k$ for *s*. Let y_k be the k^{th} value of the variable of interest, and $\mathbf{x}_k = \left(x_{k1},..,x_{kj},..,x_{kJ}\right)'$ an auxiliary variable vector of dimension $j=(1,...,J)$, where both y_k and $\mathbf{x}_k$ are observed for the sample elements $k\in s$. The Horvitz-Thompson estimator for the total of *y* is $\hat{Y}_{HT} = \sum_s d_k y_k$. In a complete response situation, with known auxiliary totals for the $j=(1,..,J)$ auxiliary variables, $\mathbf{X}=\left(X_1,..,X_j,..,X_J\right)' = \left(\sum_U x_{k1},..,\sum_U x_{kj},..,\sum_U x_{kJ}\right)'$, Deville and Särndal (1992) defined calibration as a method to find the calibrated weights w_k which minimize a distance measure $G\left(w_k,d_k\right)$ between the calibrated weights w_k and the base weights d_k. This minimization of the distance function is subject to the constraint that the calibration-weighted total of the auxiliary variable values $\sum_s w_k x_{kj}$ equals the known population total for the auxiliary $X_j, \forall j=1,\ldots,J$ as

$$\sum_s w_k \mathbf{x}_k = \mathbf{X} \qquad (2.1)$$

This calibration problem results in final calibrated weights

$$w_k = d_k F\left(q_k \mathbf{x}'_k \lambda\right) \tag{2.2}$$

where $F\left(q_k \mathbf{x}'_k \lambda\right)$ is the inverse of $\partial G\left(w_k, d_k\right) / \partial w_k$. λ denotes a vector of Lagrange multipliers used in the minimization and q_k is a positive value which scales the calibrated weights in (2.2). It is common practice to use $q_k = 1$. As an alternative to the distance minimization approach, Estevao and Särndal (2000) introduced the functional form approach to build the calibration estimators. Since both approaches lead to the same estimators, we will focus on the first one.

Many distance measures have been proposed for calibration. Deville and Särndal (1992) defined the desirable properties of these functions as (1) for every fixed $d_k > 0$, $G\left(w_k, d_k\right)$ is nonnegative, differentiable with respect to w_k , strictly convex and $G\left(d_k, d_k\right) = 0$, and (2) $\partial G\left(w_k, d_k\right) / \partial w_k$ is continuous. Empirically, there are small differences in the calibrated estimates derived from different distance measures (Singh & Mohl, 1996; Stukel, Hidiroglou, & Särndal, 1996). We will focus here on the linear case in which the chi-square distance function $\left(w_k - d_k\right)^2 / 2d_k$ is used, and $q_k = 1$ is assumed.

Under the chi-square distance function, calibration solution finds $w_k, k \in s$ by minimizing the distance function $\sum_s \left(w_k - d_k^*\right)^2 / 2d_k^*$ subject to the calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$, where d_k^* are arbitrary initial weights (a base weight or an adjusted version). This minimization generates the Lagrange multiplier vector

$$\lambda' = \left(\sum_U \mathbf{x}_k - \sum_s d_k^* \mathbf{x}_k\right)' \left(\sum_s d_k^* \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \tag{2.3}$$

When the calibration factor is $g_k = (1 + \lambda' \mathbf{x}_k)$, the final calibrated weights is

$$w_k = d_k^* (1 + \lambda' \mathbf{x}_k) = d_k^* \left[1 + \left(\sum_U \mathbf{x}_k - \sum_s d_k^* \mathbf{x}_k\right)\left(\sum_s d_k^* \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k\right] \tag{2.4}$$

and the calibrated estimated total of *y* is

$$\hat{Y}_w = \sum_s w_k y_k \tag{2.5}$$

The auxiliary variables' vector characterizes the final calibration estimator. Let E_ξ and V_ξ denote the expectation and variance with respect to the calibration model ξ. Under the univariate auxiliary variable $\mathbf{x}_k = 1$ for all $k \in U$, corresponding to the following *common mean model*

$$\begin{cases} E_\xi(y_k) = \beta \\ V_\xi(y_k) = \sigma^2 \end{cases} \tag{2.6}$$

when the overall population total is $\mathbf{X} = N$, the calibration factor is $g_k = N\left(\sum_s d_k^*\right)^{-1}$, and the calibrated estimated total of *y* is the well-known expansion estimator

$$\hat{Y}_w = N \sum_s d_k^* y_k \left(\sum_s d_k^*\right)^{-1} \tag{2.7}$$

When $\mathbf{x}_k = x_k$ for all $k \in U$, which corresponds to *the ratio model* (y_k / x_k is constant on average for any fixed x_k), the expectation and variance of y_k are

$$\begin{cases} E_{\xi}\left(y_k\right)=\beta x_k \\ V_{\xi}\left(y_k\right)=\sigma^2 x_k \end{cases} \tag{2.8}$$

When $\mathbf{X}=X$, the calibration factor is $g_k = X\left(\sum_s d_k^* x_k\right)^{-1}$, and the calibrated estimated total of *y* is the well-known ratio estimator

$$\hat{Y}_w = X\sum_s d_k^* y_k \left(\sum_s d_k^* x_k\right)^{-1} \tag{2.9}$$

For the multivariate auxiliary variable $\mathbf{x}_k = \left(1, x_k\right)$ for all $k \in U$, corresponding to *the simple regression model with an intercept*, assume the same model for all elements,

$$\begin{cases} E_{\xi}\left(y_k\right)=\alpha+\beta x_k \\ \quad V_{\xi}\left(y_k\right)=\sigma^2 \end{cases} \tag{2.10}$$

Here x_k is the value for element *k* of a continuous variable *x*, and the population total vector is $\mathbf{X} = \left(N, X\right)$. The calibrated estimated total of *y* is the well-known regression estimator

$$\hat{Y}_w = \sum_s w_k y_k = \hat{Y}_{HT} + \left(\sum_U x_k - \sum_s d_k^* x_k\right)\hat{B}_s \tag{2.11}$$

where $\hat{Y}_{HT}$ is weighted by the modified base weights d_k^*, and $\hat{B}_s = \sum_s d_k^* \mathrm{x}_k y_k \left(\sum_s d_k^* \mathrm{x}_k \mathrm{x}_k'\right)^{-1}$. All the previous estimators, in (2.7), (2.9) and (2.11), are special cases of the well-known Generalized Regression Estimator (GREG), which is a general estimator derived from the chi-square distance function. At the same time, Deville and Särndal (1992) found that calibration estimators derived with distance functions other than the chi-square are asymptotically equivalent to the GREG. Therefore, the GREG variance estimator can be used interchangeably for any of the other calibration estimators.

The variance of $\hat{Y}_w$ is

$$V\left(\hat{Y}_w\right)=\sum\sum\nolimits_U\left(\pi_{kl}-\pi_k\pi_l\right)\left(\frac{e_k}{\pi_k}\right)\left(\frac{e_l}{\pi_l}\right) \tag{2.12}$$

where $\pi_{kl}=p\left(k \,\&\, l\in s\right)$, $\pi_l=p\left(l\in s\right)$, $e_k=y_k-\mathbf{x}_k'\mathrm{B}_U$, and B_U can be written as $\mathrm{B}_U=\sum\nolimits_U \mathbf{x}_k y_k\left(\sum\nolimits_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}$. The corresponding estimated variance is

$$\hat{v}\left(\hat{Y}_w\right)=\sum\sum\nolimits_s\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\left(w_k\hat{e}_k\right)\left(w_l\hat{e}_l\right) \tag{2.13}$$

where $\hat{e}_k=y_k-\mathbf{x}_k'\hat{\mathrm{B}}_{ws}$ and $\hat{\mathrm{B}}_{ws}=\sum\nolimits_s w_k\mathbf{x}_k y_k\left(\sum\nolimits_s w_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}$.

Generally, the idea behind calibration appeals to practitioners since the GREG estimator in (2.11) can be written as a linear combinations of observations y_k with calibrated weights w_k that are sample-dependent (Deville & Sarndal, 1992). At the same time, the approach assures the external consistency, where the estimated totals of auxiliary variables are the same as the population totals.

In the following sections, the implicit potential of calibration method for combining data will be explored. Since the main idea behind calibration is to find a set of weights which guarantee that the estimated auxiliary totals conform to known population totals, the same idea can be used to combine two samples. As opposed to the standard dual frame estimators discussed in Chapter 1, the Joint Calibration approach will be introduced as a method for combining dual frame samples. In addition to the practical simplicity of Joint Calibration, it is flexible enough to accommodate varying forms of available auxiliary variables in dual frame estimation.

In single sample designs, a strong correlation between auxiliary variables and the study variable implies that the weights that perform well for the auxiliary variables should also perform well for the study variable and results in asymptotically unbiased calibrated estimates (Deville & Sarndal, 1992). The same idea can be applied to the dual frame estimation, where strong associations between the auxiliary variables and the study variable results in asymptotically unbiased dual frame estimates, as proved in Proposition 2.1 in Section 2.5.

2.3 Joint Calibration Estimator (JCE)

Under the dual frame design, where $E(\ )$ denotes design-based expectation, $E\left(\sum_{s_A} d_k \mathbf{x}_k\right) = \mathbf{X}_A$, $E\left(\sum_{s_B} d_k \mathbf{x}_k\right) = \mathbf{X}_B$ and $E\left(\sum_{s_A} d_k \mathbf{x}_k + \sum_{s_B} d_k \mathbf{x}_k\right) \neq \mathbf{X}$, calibration conditioning on $\sum_{s_A} w_k \mathbf{x}_k + \sum_{s_B} w_k \mathbf{x}_k = \mathbf{X}$ should achieve $E\left(\sum_{s_A} w_k \mathbf{x}_k + \sum_{s_B} w_k \mathbf{x}_k\right) = \mathbf{X}$. Consequently, a powerful set of auxiliary variables, that are strong predictors for the study variable y, should result in $E\left(\sum_{s_A} w_k y_k + \sum_{s_B} w_k y_k\right) \simeq \mathrm{Y}$, as proven in Proposition 2.1 in Section 2.5.

Under the complete response assumption, calibrated estimates can be parameterized under the dual frame design through re-deriving the calibration factors as explicit components for each sample of the dual frame sample. By jointly calibrating the two datasets, the calibration problem will be to find final weights w_k, $k \in s$ to satisfy the calibration equation

$$\sum\nolimits_{s} w_k \mathbf{x}_k = \sum\nolimits_{s_A} w_k \mathbf{x}_k + \sum\nolimits_{s_B} w_k \mathbf{x}_k = \mathbf{X} \tag{2.14}$$

through minimizing the following distance function

$$\sum\nolimits_{s} \left(w_k - d_k\right)^2 \Big/ 2d_k \tag{2.15}$$

The distance function can be split into two components for the two samples A and B as follows

$$\sum\nolimits_{s_A} \left(w_k - d_k\right)^2 \Big/ 2d_k + \sum\nolimits_{s_B} \left(w_k - d_k\right)^2 \Big/ 2d_k \tag{2.16}$$

Consequently, the joint calibration weights are

$$w_k = \begin{cases} d_k \left(1 + \lambda' \mathbf{x}_k\right) & k \in s_A \\ d_k \left(1 + \lambda' \mathbf{x}_k\right) & k \in s_B \end{cases} \tag{2.17}$$

where $\lambda' = \left(\sum\nolimits_U \mathbf{x}_\text{k} - \sum\nolimits_s d_k \mathbf{x}_\text{k}\right)' \left(\sum\nolimits_s d_k \mathbf{x}_\text{k} \mathbf{x}_\text{k}'\right)^{-1}$ and the joint calibration factor is $g_k = \left(1 + \lambda' \mathbf{x}_k\right)$.

Recall, the dual frame estimator for the total of *y*, can be written as in equation (1.3), and the weighted version expressed as in equations (1.4) and (1.5). A modified version of equation (1.4) can be written as

$$\hat{Y} = \sum\nolimits_{s_a} d_k y_k + \sum\nolimits_{s_{ab}^A} m_k d_k y_k + \sum\nolimits_{s_{ab}^B} m_k d_k y_k + \sum\nolimits_{s_b} d_k y_k \tag{2.18}$$

Consequently, where auxiliary variable $\mathbf{x}_k = 1$ for $k \in U$, the main constraint in (2.14) can be written as

$$\sum\nolimits_{s} w_k = N \tag{2.19}$$

and the following constraints can be added to the calibration minimization problem

$$w_k = d_k \ \forall \ k \in s_a \cup s_b \tag{2.20}$$

This constraint is identical to

$$\sum_{s_a} w_k = \sum_{s_a} d_k^* = N_a \tag{2.21}$$

and

$$\sum_{s_b} w_k = \sum_{s_b} d_k^* = N_b \tag{2.22}$$

In (2.21), $d_k^* = \left(N_a \middle/ \sum_{s_a} d_k \right) d_k$, whereas, in (2.22), $d_k^* = \left(N_b \middle/ \sum_{s_b} d_k \right) d_k$. In fact, the calibration problem with the three constraints, (2.19), (2.21) and (2.22), is identical to post-stratifying the sample by the design domain totals N_a, N_{ab} and N_b. Therefore, calibrating by these totals yields the unbiased dual frame estimator (1.3), where the modification factors m_k^A and m_k^B for the overlap domain have the same value m_k and can be written as

$$m_k = \ N_{ab} \middle/ \left(\sum_{s_{ab}^A} d_k + \sum_{s_{ab}^B} d_k \right) \ \forall k \in s_{ab}^A \cup s_{ab}^B \tag{2.23}$$

At the same time, the joint calibration factor can be written as

$$g_k = \begin{cases} N_a \middle/ \sum_{s_a} d_k & k \in s_a \\ N_{ab} \middle/ \left(\sum_{s_{ab}^A} d_k + \sum_{s_{ab}^B} d_k \right) & k \in s_{ab}^A \cup s_{ab}^B \\ N_b \middle/ \sum_{s_b} d_k & k \in s_b \end{cases} \tag{2.24}$$

It is worth noting that by using the joint calibration factor in (2.24), this version of JCE is identical to the post-stratified version of the Fixed Weight Estimator (FWE)

$$\hat{Y}_{FWE}^{post} = \frac{N_a}{\hat{N}_a}\hat{Y}_a + \frac{N_{ab}}{\hat{N}_{ab}}\left(\theta\hat{Y}_{ab}^{A} + (1-\theta)\hat{Y}_{ab}^{B}\right) + \frac{N_b}{\hat{N}_b}\hat{Y}_b \tag{2.25}$$

where $\theta = .5$ and $\hat{N}_{ab} = \left(\theta\hat{N}_{ab}^{A} + (1-\theta)\hat{N}_{ab}^{B}\right)$. Interestingly, the same conclusion can be reached based on Remark 2.1 proved by Lündstrom and Sarndal (1999):

"Suppose the population is divided into P groups $U_1,...,U_p,....,U_P$ and that the group total $\sum_{U_p} \mathrm{x}_k$ is known for $p = 1,...,P$, and used in the calibration. Let c_p, $p = 1,...,P$, be arbitrary positive constants. Then the initial weights $d_k^* = d_k$ and the initial weights $d_k^* = c_p d_k$,for $k \in r_p$, give exactly the same final weights when $\sum_s \left(w_k - d_k^*\right)^2 / d_k^* q_k$ is minimized and $q_k = 1/\mu' \mathrm{x}_k$ holds."

Under multiple frame design, when the population is divided into *P* domains and when d_k^* is replaced by $c_p d_k^*$ in the distance function, where c_p is the dual frame composite factor, minimizing the chi-square distance

$$\sum_P \sum_{s_p} \frac{\left(w_k - c_p d_k^*\right)^2}{2c_p d_k^*} \tag{2.26}$$

under the constraints $\sum_{s_p} w_k = N_p$, achieves an asymptotically unbiased calibration estimator, and implicitly combines the samples. Under dual frame design, c_p can be written as

$$c_p = \begin{cases} 1 & p \in s_a \\ \theta & p \in s_{ab}^A \\ 1-\theta & p \in s_{ab}^B \\ 1 & p \in s_b \end{cases} \tag{2.27}$$

However, we should have a fixed c_p within each domain. So using $\theta = 0.5$, which follows the fixed weights dual frame estimator (FWE), results in the asymptotically unbiased calibrated estimator. In this case, considering that N, N_A, N_B and N_{ab} are known, the calibrated total of y can be written in form of domain-level means as

$$\hat{Y} = \frac{N - N_B}{\hat{N}_a}\hat{Y}_a^A + \frac{N - N_A}{\hat{N}_b}\hat{Y}_b^B + \frac{N_{ab}}{\theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B}\left(\theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B\right) \tag{2.28}$$

When N_{ab} is unknown, the calibrated total of y can be written as

$$\hat{Y} = \frac{N - N_B}{\hat{N}_a}\hat{Y}_a^A + \frac{N - N_A}{\hat{N}_b}\hat{Y}_b^B + \frac{N_A + N_B - N}{\theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B}\left(\theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B\right) \tag{2.29}$$

Consequently, calibrating the dual frame to the totals of domains (*a, ab & b*) is identical to the FWE and gives unbiased estimates. This means that the calibration constraints or the auxiliary variables used determine forms of the JCE, some of which might be identical to the standard dual frame estimators. In the next section, more general forms of JCE are discussed.

The JCE can be applied to the general case of multiple frames. Under the dual frame design, the JCE for population total of y can be written as

$$\hat{Y}_{JCE} = \sum\nolimits_s w_k y_k = \sum\nolimits_{s_A} w_k y_k + \sum\nolimits_{s_B} w_k y_k \tag{2.30}$$

where $w_k = d_k \left(1+\lambda'\mathbf{x}_k\right)$ and $\lambda' = \left(\sum_U \mathbf{x}_\mathrm{k} - \sum_{s_A} d_k\mathbf{x}_\mathrm{k} - \sum_{s_B} d_k\mathbf{x}_\mathrm{k}\right)' \left(\sum_{s_A} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k} + \sum_{s_B} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k}\right)^{-1}$.

Under multiple frame designs, when the population is divided into *P* domains, the JCE for population total of *y* can be written as

$$\hat{Y}_{JCE} = \sum_P \sum_{s_p} w_k y_k \tag{2.31}$$

where $w_k = d_k \left(1+\lambda'\mathbf{x}_k\right)$ and λ' can be written as

$$\lambda' = \left(\sum_U \mathbf{x}_\mathrm{k} - \sum_P \sum_{s_p} d_k\mathbf{x}_\mathrm{k}\right)' \left(\sum_P \sum_{s_p} d_k\mathbf{x}_\mathrm{k}\mathbf{x}'_\mathrm{k}\right)^{-1} \tag{2.32}$$

2.4 Examples of Joint Calibration Estimators

The auxiliary variable vector characterizes the final JCE for dual frame estimation. For example, under the univariate auxiliary variable $\mathbf{x}_k = 1$ for $k \in U$, which corresponds to *the common mean model* in (2.6), where the overall population total is $\mathbf{X} = N$, the joint calibration factor is

$$g_k = N\left(\sum_{s_A} d_k + \sum_{s_B} d_k\right)^{-1} \tag{2.33}$$

By calibrating the concatenated or "stacked" datasets for each frame's sample, $\sum_{s_A} w_k\mathbf{x}_k + \sum_{s_B} w_k\mathbf{x}_k = N$. This JCE estimate is appropriate when it is thought that the true common mean β is the same for all $k \in U$. However, another JCE estimate is appropriate when it is thought that β vary between design domains $d=(a,ab,b)$. This estimate uses the calibration factor in (2.24).

When $\mathbf{x}_k = x_k$ for $k \in U$, corresponding to *the ratio model* in (2.8), and when $\mathbf{X} = X$, the joint calibration factor is

$$g_k = X\left(\sum\nolimits_{s_A} d_k x_k + \sum\nolimits_{s_B} d_k x_k\right)^{-1} \tag{2.34}$$

By calibrating the stacked dataset, $\sum_{s_A} w_k \mathbf{x}_k + \sum_{s_B} w_k \mathbf{x}_k = X$. This JCE estimate is appropriate when it is thought that βx_k is the same, for all $k \in U$.

Another JCE estimate is appropriate when it is thought that βx_k vary between design domains $d=(a,ab,b)$. This estimate uses the calibration factor

$$g_k = \begin{cases} X_a \Big/ \sum_{s_a} d_k x_k & k \in s_a \\ X_{ab} \Big/ \left(\sum_{s_{ab}^A} d_k x_k + \sum_{s_{ab}^B} d_k x_k\right) & k \in s_{ab}^A \cup s_{ab}^B \\ X_b \Big/ \sum_{s_b} d_k x_k & k \in s_b \end{cases} \tag{2.35}$$

Obviously, this estimate requires knowledge of the separate totals (X_a, X_{ab}, X_b).

Under the multivariate auxiliary variable $\mathbf{x}_k = (1, x_k)$ for $k \in U$, which corresponds to *the simple regression model with intercept* in (2.10), where the population total vector is $\mathbf{X} = (N, X)$ the calibrated estimate of the total, $\hat{Y}_{JCE}$, can be written as

$$\hat{Y}_{JCE} = \hat{Y}_{HT}^{A} + \hat{Y}_{HT}^{B} + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{s_A} d_k x_k + \sum\nolimits_{s_B} d_k x_k \right) \right) \hat{B}_s^{A,B} \tag{2.36}$$

where $\hat{B}_s^{A,B} = \left(\sum\nolimits_{s_A} d_k \mathbf{x}_k y_k + \sum\nolimits_{s_B} d_k \mathbf{x}_k y_k \right) \left(\sum\nolimits_{s_A} d_k \mathbf{x}_k \mathbf{x}_k' + \sum\nolimits_{s_B} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1}$. This JCE estimate is appropriate when it is thought that $\alpha + \beta x_k$ is the same, for all $k \in U$. With more than one auxiliary variable, the multivariate formula can be written as

$$\hat{Y}_{JCE} = \hat{Y}_{HT}^{A} + \hat{Y}_{HT}^{B} + \left(\sum\nolimits_{U} \mathbf{x}_k - \left(\sum\nolimits_{s_A} d_k \mathbf{x}_k + \sum\nolimits_{s_B} d_k \mathbf{x}_k \right) \right) \hat{B}_s^{A,B} \tag{2.37}$$

where $\mathbf{x}_k = \left(x_{k1}, .., x_{kj}, .., x_{kJ} \right)'$ is the auxiliary variable vector with $j = (1, ..., J)$ dimensions. Interestingly, since $\left(\sum\nolimits_{s_A} d_k \mathbf{x}_k + \sum\nolimits_{s_B} d_k \mathbf{x}_k \right)$ is always greater than $\sum\nolimits_{U} \mathbf{x}_k$, the term $\left(\sum\nolimits_{U} \mathbf{x}_k - \left(\sum\nolimits_{s_A} d_k \mathbf{x}_k + \sum\nolimits_{s_B} d_k \mathbf{x}_k \right) \right) \hat{B}_s^{A,B}$ in (2.37) can be viewed as a negative sign correction factor applied to the biased summation of the $\hat{Y}_{HT}^{A}$ and $\hat{Y}_{HT}^{B}$ from both samples. Note that all the JCE forms can be derived from the general JCE form in (2.37).

Another interesting multivariate calibration estimator is the complete post-stratified estimator, which corresponds to *the group mean model*, calibrating on known post-stratified cell counts. When the sizes of the population groups N_p and the classification vector used to code membership in one of *P* mutually exclusive and exhaustive groups are known, and $\mathbf{x}_k = \gamma_k = \left(\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk} \right)'$ is the auxiliary variables vector, the calibrated estimator is the well-known post-stratified estimator. The joint calibration factor takes the form

$N_p \Big/ \left(\sum_{s_p^A} d_k + \sum_{s_p^B} d_k \right)$ where s_p^A denotes the sample cell $U_p \cap s_A$ and s_p^B denotes the sample cell $U_p \cap s_B$. The calibrated estimator of the total of the study variable *y* can be written as

$$\hat{Y}_{JCE} = \sum_P \frac{N_p}{\left(\sum_{s_p^A} d_k + \sum_{s_p^B} d_k \right)} \left(\sum_{s_p^A} d_k y_k + \sum_{s_p^B} d_k y_k \right) \tag{2.38}$$

In *the group mean model*, it is implicitly assumed that mean and variance are shared by all elements within the same group *p* as

$$\begin{cases} E(y_k) = \beta_p \\ V(y_k) = \sigma_p^2 \end{cases} \tag{2.39}$$

Similarly, when the group totals X_p are known and $\mathbf{x}_k = x_k \gamma_k = \left(x_{1k}\gamma_{1k}, \ldots, x_{pk}\gamma_{pk}, \ldots, x_{Pk}\gamma_{Pk} \right)'$ is used as the auxiliary variables vector, this corresponds to *the group ratio model*, where mean and variance are shared by all elements within the same group *p* as

$$\begin{cases} E(y_k) = \beta_p x_k \\ V(y_k) = \sigma_p^2 x_k \end{cases} \tag{2.40}$$

Both *the group mean model* and *the group ratio model* may be classified under *the group models* (Sarndal, Swensson & Wretman, 1992). Since the groups in the group models can serve as strata, JCE would have better performance if this informative design has been incorporated into the auxiliary variable vector.

2.5 Joint Calibration Estimator Bias

In this section, an approximate JCE bias is derived. This helps in understanding the mechanism of the joint calibration approach in combining the dual frame design samples in contrast to the other dual frame estimators. At the same time, it highlights the JCE as a model-assisted design-based estimator for which the design-based bias properties are affected by the association between the study variable y and the auxiliary variable vector $\mathbf{x}$.

Proposition 2.1

The bias of the JCE estimator, $\hat{Y}_{JCE}$, in (2.30) is given approximately by

$$Bias\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{U_{ab}} e_k^{A,B} \tag{2.41}$$

where

$$e_k^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_U^{A,B}\right), \qquad \mathrm{B}_U^{A,B} = \left(\sum\nolimits_{U_A} \mathbf{x}_k \mathbf{x}_k' + \sum\nolimits_{U_B} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum\nolimits_{U_A} \mathbf{x}_k y_k + \sum\nolimits_{U_B} \mathbf{x}_k y_k\right) \tag{2.42}$$

Appendix 2.1 presents the derivation of (2.41).

Note that the dual frame estimation bias can be derived from (1.2) in Chapter 1 as

$$Bias\left(\hat{Y}_A + \hat{Y}_B\right) = \sum\nolimits_{U_{ab}} y_k \tag{2.43}$$

This means that the joint calibration approach uses $\mathbf{x}_k' \mathrm{B}_U^{A,B}$ to attenuate the bias for each $k \in U_{ab}$ to reduce the bias in (2.43). Therefore, the reduction in dual frame estimation bias due to the joint calibration is $\sum\nolimits_{U_{ab}} \mathbf{x}_k' \mathrm{B}_U^{A,B}$, which is the difference between (2.43) and (2.41). Proposition

2.1 emphasizes the need to identify powerful auxiliary variables that can predict study variable *y*. The more $\mathbf{x}_k' \mathrm{B}_U^{A,B}$ is able to predict y_k for each $k \in U_{ab}$, the more reduction in bias. The bias of $\hat{Y}_{JCE}$ in (2.41) is independent of the sampling design used to draw s_A and s_B as long as the set of auxiliary variables $\mathbf{x}_k$ is the same.

Corollary 2.1

When a perfect linear relationship exists in the population between the study variable y_k and the auxiliary vector $\mathbf{x}_k$, as $y_k = \mathbf{x}_k' \mathrm{B}_U$, for every $k \in U$, the bias of the JCE estimator in (2.41) can be written as

$$Bias\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{U_{ab}} \mathbf{x}_k' \left(\mathrm{B}_U - \mathrm{B}_U^{A,B}\right) = 0 \qquad (2.44)$$

This is due the fact that when this perfect linear relationship between y_k and x_k exists, $\mathrm{B}_U^{A,B} = \mathrm{B}_U$. That is, the bias of $\hat{Y}_{JCE}$ is a function of the difference between two regression vectors $\mathrm{B}_U^{A,B}$ and B_U. This perfect relationship will not hold in practice. However, the bias in (2.41) will be reduced if the perfect linear relationship between y_k and $\mathbf{x}_k$ comes close to being attained. We should use auxiliary variables $\mathbf{x}_k$ such that the residuals $e_k^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_U^{A,B}\right)$ are small. This happens when the residuals $e_k = \left(y_k - \mathbf{x}_k' \mathrm{B}_U\right)$ are small. Using such a set of auxiliary variables $\mathbf{x}_k$ guarantees reduced bias and variance of the JCE as a dual frame estimator.

Corollary 2.1 helps us in understanding the relationship between the study variable *y* and the auxiliary variable vector **x**. The performance of the JCE is controlled by the association between y and **x**, where the best performance happens when (1) **x** more closely matches the population model or (2) **x** includes some strong correlates of y. Although the first case results in greater reduction in bias, the second case is more appealing due to an unknown population model required for the first case.

Assuming that the same model holds for all units in the whole population, $\frac{1}{N_{ab}}\sum_{U_{ab}} e_k^{A,B}$ asymptotically follows $N(0,V)$ where V is $O\left(N_{ab}^{-1}\right)$, the bias of the mean estimator $Bias\left(\hat{\bar{Y}}_{JCE}\right)=\frac{1}{N}\sum_{U_{ab}} e_k^{A,B}$, where $\hat{\bar{Y}}_{JCE}=\hat{Y}_{JCE}/N$, converges in probability to 0 in large populations, $Bias\left(\hat{\bar{Y}}_{JCE}\right)\xrightarrow{p}0$. This is due to the fact that the variance of the zero mean $Bias\left(\hat{\bar{Y}}_{JCE}\right)=\frac{N_{ab}}{N}\frac{1}{N_{ab}}\sum_{U_{ab}} e_k^{A,B}$ is proportional to $P_{ab}^2 O\left(N_{ab}^{-1}\right)\approx\frac{P_{ab}}{N}$, where $\frac{N_{ab}}{N}\to P_{ab}$, and $\frac{P_{ab}}{N}\to 0$ as $N\to\infty$. This means that the JCE estimator of mean, $\hat{\bar{Y}}_{JCE}$, is a consistent estimator of population mean, $\bar{Y}$.

2.6 Joint Calibration Estimator Variance estimation

Where variance of $\hat{Y}_w$ under single frame design can be written as in (2.12), under dual frame design variance of $\hat{Y}_{JCE}$ can be written as

$$V\left(\hat{Y}_{JCE}\right)=\sum\sum_{U_A}\Delta_{kl}^{A}\left(\frac{e_k^A}{\pi_k^A}\right)\left(\frac{e_l^A}{\pi_l^A}\right)+\sum\sum_{U_B}\Delta_{kl}^{B}\left(\frac{e_k^B}{\pi_k^B}\right)\left(\frac{e_l^B}{\pi_l^B}\right)+\sum\sum_{U_{ab}}\Delta_{kl}^{ab}\left(\frac{e_k^{ab}}{\pi_k^{ab}}\right)\left(\frac{e_l^{ab}}{\pi_l^{ab}}\right) \qquad (2.45)$$

where $s_{ab} = s_A \cap s_B$, for $D = (A, B, ab)$, $\Delta_{kl}^D = \left(\pi_{kl}^D - \pi_k^D \pi_l^D\right)$, $\pi_{kl}^D = p\left(k \,\&\, l \in s_D\right)$, $\pi_k^D = p\left(k \in s_D\right)$, $\pi_l^D = p\left(l \in s_D\right)$, $e_k^D = y_k - \mathbf{x}_k' \mathrm{B}_{U_D}$, and $\mathrm{B}_{U_D} = \sum_{U_D} \mathbf{x}_k y_k \left(\sum_{U_D} \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$. Assuming negligible values of π_{kl}^{ab}, π_k^{ab} and π_l^{ab}, the corresponding estimated variance is

$$\hat{v}\left(\hat{Y}_{JCE}\right) = \sum\sum_{s_A} \frac{\Delta_{kl}^A}{\pi_{kl}} \left(w_k \hat{e}_k^A\right)\left(w_l \hat{e}_l^A\right) + \sum\sum_{s_B} \frac{\Delta_{kl}^B}{\pi_{kl}} \left(w_k \hat{e}_k^B\right)\left(w_l \hat{e}_l^B\right) \tag{2.46}$$

where $\hat{e}_k^D = y_k - \mathbf{x}_k' \hat{\mathrm{B}}_{ws_D}$, and $\hat{\mathrm{B}}_{ws_D} = \sum_{s_D} w_k \mathbf{x}_k y_k \left(\sum_{s_D} w_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$.

2.7 Simulation study

Simulation studies were used to evaluate the performance of the JCE relative to the standard FWE dual frame estimator under the complete response assumption. In these studies, the estimation bias and mean squared error are used to compare different estimators. The simulation studies focus on the estimate of the population total of a variable *y*.

Two population models were used to generate simulated populations. For both models, the finite population size was $N = 100{,}000$ with domains population sizes $N_a = 40{,}000$, $N_{ab} = 50{,}000$ and $N_b = 10{,}000$. The population was grouped into *J=6* strata with sizes $N_1 = 10{,}000$, $N_2 = 20{,}000$, $N_3 = 30{,}000$, $N_4 = 25{,}000$, $N_5 = 5{,}000$ and $N_6 = 10{,}000$. Frame sizes are $N_A = 90{,}000$ (all cases in domains *a* and *ab*) and $N_B = 60{,}000$ (all cases in domains *ab* and *b*). The distribution of the population elements over the strata and the domains is presented in Table 2.1.

Table 2.1: Distribution of the population elements over the six strata and the three domains.

	Frames and domains			
	A			
		B		
Strata	*a*	*ab*	*b*	Total
1	10,000			10,000
2	20,000			20,000
3	10,000	20,000		30,000
4		25,000		25,000
5		5,000		5,000
6			10,000	10,000
Total	40,000	50,000	10,000	100,000

The first population model is a *common linear regression model* (CLR), $y_{jk} = x_{jk} + \varepsilon_{jk}$, for $k = 1,..,N$ and $j = 1,...,6$ strata, where $x_{jk} \sim N(\mu_x, \sigma_x)$ and $\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$. Here the mean of *y* is the same for all population strata and design domains. The second population model is a *group linear regression model* (GLR), which can be written as the first model but with $x_{jk} \sim N(\mu_{xj}, \sigma_x)$ and $\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$. In both models, an auxiliary variable, z_{dk}, was generated as $z_{dk} = \beta_o + \beta_d + \varepsilon_{dk}$, for $d = (\text{a,ab,b})$ where $\beta_o = 200$ and $\varepsilon_{dk} \sim N(0,350)$. For both the first and the second models, the simulation factors were as follows:

1. Sampling Designs
   a) Simple Sampling Design: simple random samples were selected from both frames.
   b) Complex Sampling Design: stratified sample with equal allocation to five strata from frame A, and a simple random sample from frame B.
2. Sample size
   a) Equal allocation where $n_A = 500$ and $n_B = 500$.

b) Proportional allocation where $n_A = 600$ and $n_B = 400$.

c) Extreme allocation where $n_A = 900$ and $n_B = 100$.

3. Domain means

a) Small-differences in domain means where $\beta_a = 5$, $\beta_{ab} = 6$ and $\beta_b = 7$.

b) Frame-different means where $\beta_a = 5$, $\beta_{ab} = 5$ and $\beta_b = 10$.

c) Large-differences in domain means where $\beta_a = 5$, $\beta_{ab} = 10$ and $\beta_b = 15$.

4. Correlation between y_{jk} and x_{jk}

a) The population correlation coefficient is $\rho_{xy} = 0.40$.

b) The population correlation coefficient is $\rho_{xy} = 0.60$.

c) The population correlation coefficient is $\rho_{xy} = 0.80$.

The correlation levels in the last factor determine the population model parameters as presented in Table 2.2. Regarding to the CLR model, different values of σ_x and σ_ε are deliberately assumed to generate different correlation levels. Since μ_{xj} does not contribute to the correlation, it is almost fixed across the correlation levels. This applies for the GLR model, except that μ_{xj} is different across the 6 strata.

Table 2.2: Model parameters based on correlation levels between y_{jk} and x_{jk}.

	ρ_{xy}		
Model parameters	$\rho_{xy} = 0.40$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.80$
CLR Model			
$x_{jk} \sim N(\mu_x, \sigma_x)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$
GLR Model			
$x_{1k} \sim N(\mu_{x1}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(480,384)$
$x_{2k} \sim N(\mu_{x2}, \sigma_x)$	$N(618,192)$	$N(640,288)$	$N(620,384)$
$x_{3k} \sim N(\mu_{x3}, \sigma_x)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$x_{4k} \sim N(\mu_{x4}, \sigma_x)$	$N(881,192)$	$N(919,288)$	$N(900,384)$
$x_{5k} \sim N(\mu_{x5}, \sigma_x)$	$N(1013,192)$	$N(1059,288)$	$N(1039,384)$
$x_{6k} \sim N(\mu_{x6}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(479,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$

These sets of simulation factors combine to form 108 simulation studies, 54 simulation studies for each population model. One thousand replicates of initial samples of 1,000 cases were run for each study, resulting in standard error less than 60 for difference in the biases between FWE and JCE estimators. To simulate a dual frame design, within each simulation replicate, two samples were independently drawn from both frames A and B. These samples were 'stacked' to form dual frame sample *s*.

Dual frame estimation methods were then applied to each simulated dual frame sample. FWE with $\theta = 0.5$ represents the standard fixed weight dual frame estimator, $\hat{Y}_{FWE}$. FWE with $\theta = 0.5$ means that after applying the base weights, d_k^A and d_k^B, for each sample, the combination step adjusts the base weights using a composite factor $\theta = 0.5$. The auxiliary variables were then used to calibrate the adjusted base weights to the auxiliary totals for three combinations of *x* and *z* resulting in the calibrated versions $\hat{Y}_{FWE.z}^{cal}$, $\hat{Y}_{FWE.x}^{cal}$ and $\hat{Y}_{FWE.xz}^{cal}$. Additionally, in conjunction with *x*, the design domain, D = (*a, ab, b*), and frame identifiers, F = (*A*,*B*), were used to calibrate the adjusted base weights resulting in $\hat{Y}_{FWE.xD}^{cal}$ and $\hat{Y}_{FWE.xF}^{cal}$.

For the JCE, the base weights, d_k^A and d_k^B were applied for each sample, and then the auxiliary variables *x* and *z* were used to calibrate the base weights directly, resulting in the JCE estimators, $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Additionally, under the GLR model, $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ were produced using the same auxiliary variables used in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$, respectively, except that stratum totals were also included in the calibration auxiliary variable set. Also, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xF}$ were produced using the same auxiliary variables used in $\hat{Y}_{FWE.xD}^{cal}$ and $\hat{Y}_{FWE.xF}^{cal}$, respectively.

The biases in the JCE estimates and the FWE estimates were assessed through a comparison of the survey estimate $\hat{Y}$, to the population parameter Y for the synthetic finite population. Relative Bias of parameter estimates (RB) was computed as

$$RB = \left(\frac{\sum_{i \in 1000} \hat{Y}_i}{1000} - Y \right) \times 100 \Big/ Y \tag{2.47}$$

The Relative Root Mean Squared Error (RMSE) for each estimator was computed as

$$RMSE = \sqrt{\sum_{i \in 1000} \frac{\left(\hat{Y}_i - Y\right)^2}{1000}} \times 100 \Big/ Y \tag{2.48}$$

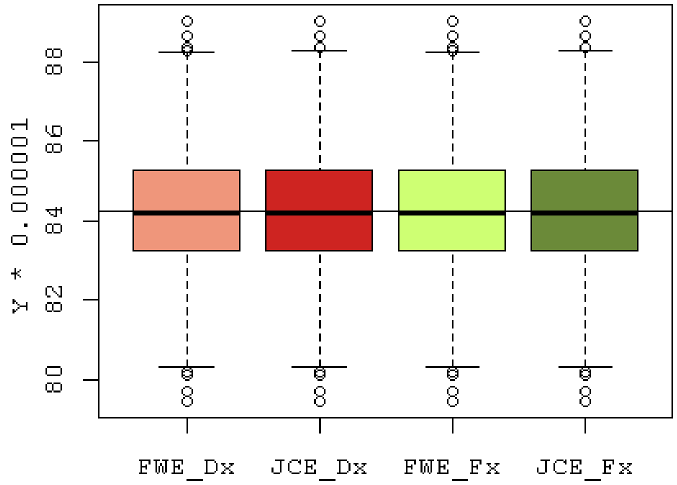
for 1000 replications or simulated dual frame samples for each simulation specification.

2.8 Simulation Results

In this section, only results for the simple sampling design are discussed, since simulation results for complex sampling designs, in Appendix 2.2, show the same patterns of results consistent with proposition 2.1. As indicated in Figure 2.1, under simple sampling design from the CLR model population, where $n_A = 500$ and $n_B = 500$, $\beta_a = 5,\ \beta_{ab} = 6$ and $\beta_b = 7$, and under 0.4 correlation between y_{jk} and x_{jk}, JCE_Dx ($\hat{Y}_{JCE.xD}$) and JCE_Fx ($\hat{Y}_{JCE.xF}$) gave exactly the same results as FWE_Dx ($\hat{Y}^{cal}_{FWE.xD}$) and FWE_Fx ($\hat{Y}^{cal}_{FWE.xF}$), respectively. This is consistent with the proof in (2.28) and (2.29) where calibrating the base weights by the population totals of design domain in $\hat{Y}_{JCE.D}$ or frame totals in $\hat{Y}_{JCE.F}$ is identical to FWE with $\theta = 0.5$, $\hat{Y}_{FWE}$. These four estimators (JCE_Dx, JCE_Fx, FWE_Dx and FWE_Fx) give almost unbiased estimates, where the estimate expectations almost equal 84,252,408, the population total Y. The same applies under all the simulation studies, as long as the design domain or the frames are considered in the calibration auxiliary variables vector. Therefore, the results for FWE_Dx and FWE_Fx are not included in the simulation tables or in our discussion.

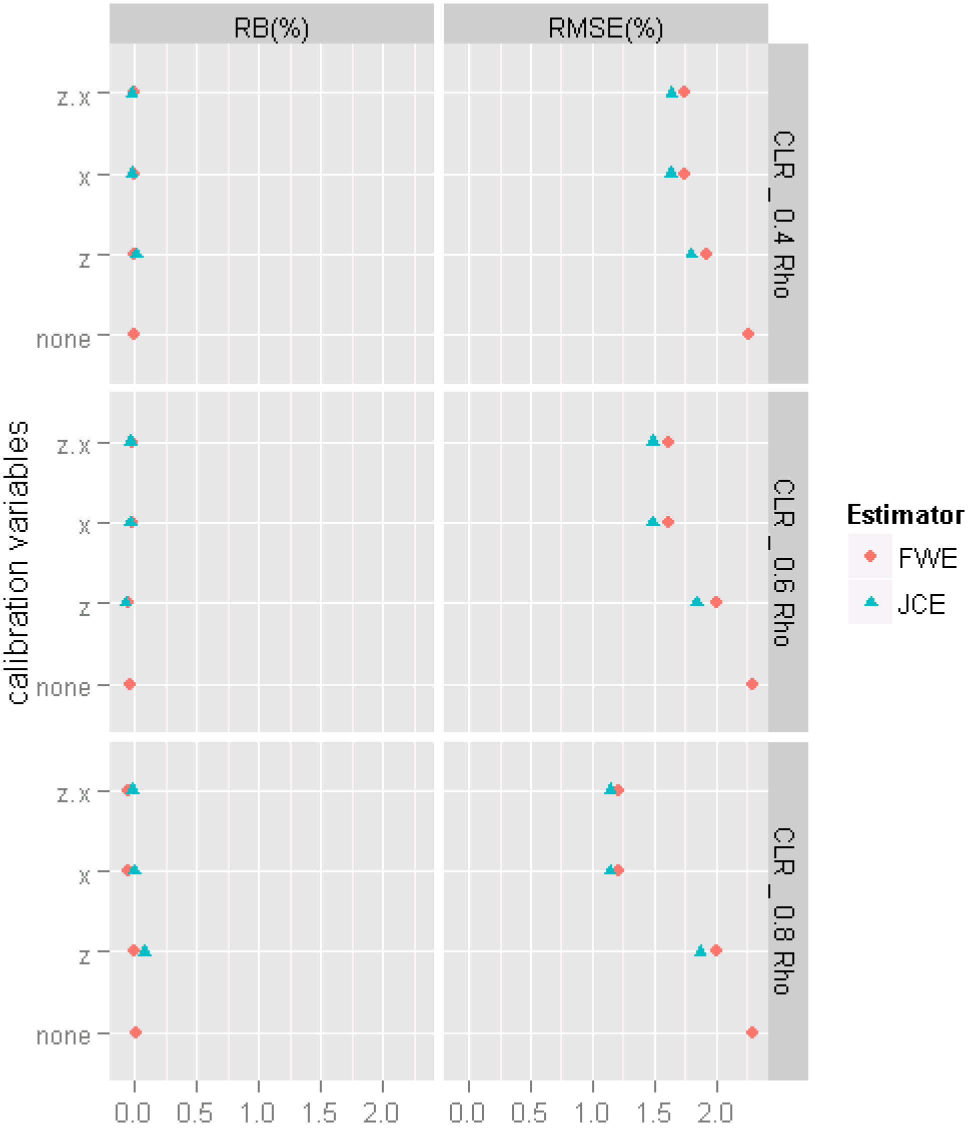
As indicated in Figures 2.2 and 2.3, under simple sampling design from the CLR or GLR model populations, where $n_A = 500$ and $n_B = 500$ and $\beta_a = 5,\ \beta_{ab} = 6$ and $\beta_b = 7$, when complete response is assumed, the standard estimator $\hat{Y}_{FWE}$ achieves nearly unbiased estimates. Under the CLR model, as indicated in Figure 2.2, the proposed $\hat{Y}_{JCE}$ estimators achieve relative biases comparable to the standard estimator $\hat{Y}_{FWE}$ or its calibrated versions, $\hat{Y}^{cal}_{FWE}$'s. This means that calibrating the 'stacked' samples directly by *z* or *x* in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ was enough to combine the two samples without the composite factor θ, under the complete response assumption. The same applies under the other sample allocations and domain means as in Table 2.3.

Figure 2.1: The standard estimator FWE and the proposed estimator JCE of $\hat{Y}$, estimated from CLR model population under simple sampling design



Under the CLR model, in Table 2.3, there are no apparent differences between the relative biases for the different simulation studies. For example, neither the association level between *y* and *x* nor the domain means have any effect on the relative biases in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ or $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.xz}$, respectively. The RMSE's, in Figure 2.3 and Table 2.4, show the same patterns as the relative biases, although RMSE's for $\hat{Y}_{JCE}$ were slightly lower than RMSE's for $\hat{Y}^{cal}_{FWE}$.

Figure 2.2: Simulation RB (%) and RMSE (%) for FWE and JCE estimators estimated under the CLR model population under simple sampling design, equal allocation and small domain mean differences

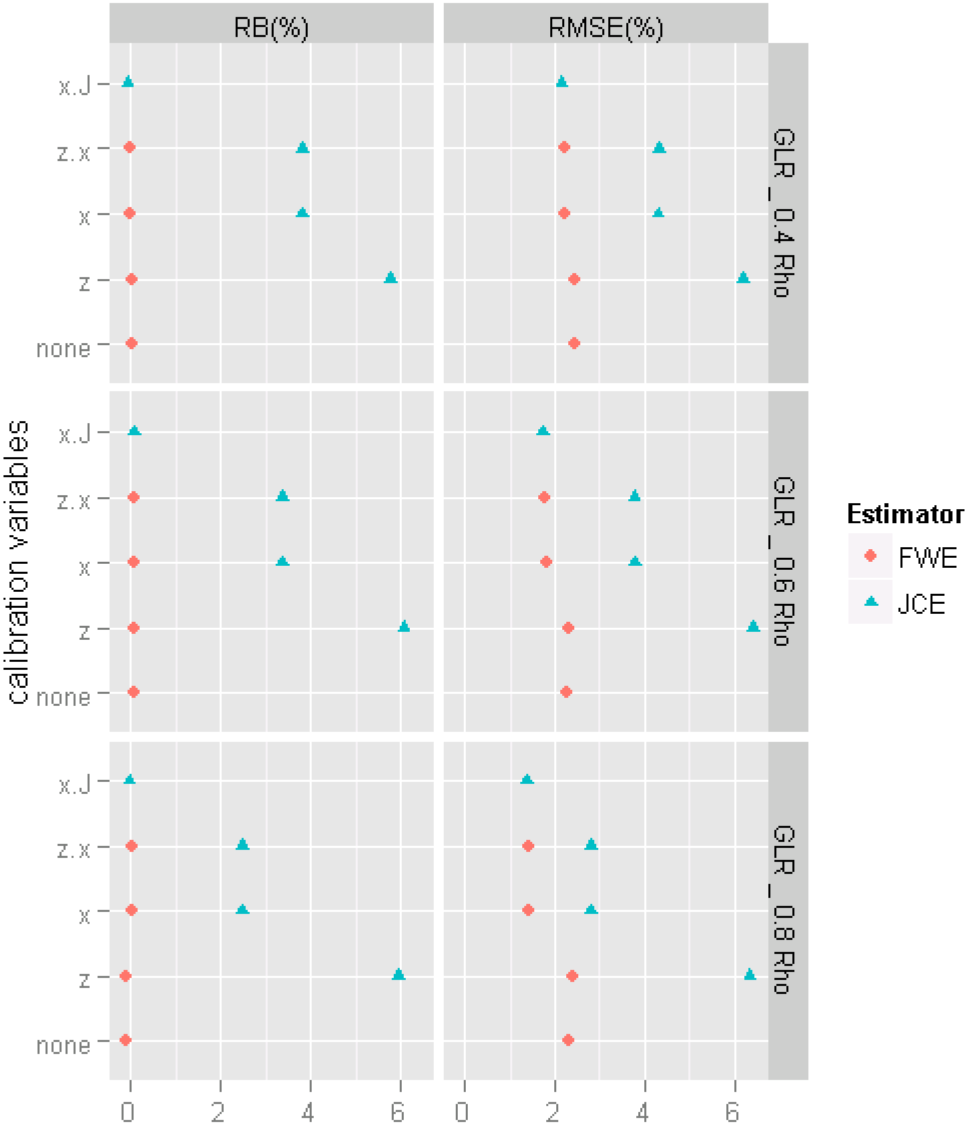


With regard to the sample allocation, in Table 2.3, there is no apparent difference between the relative biases in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ across allocations. However, the effect of the sample allocation is more obvious on the RMSE, in Table 2.4, due to the effect of the sample allocation on the estimated variance. The proportional allocation, $n_A = 600$ and $n_B = 400$, tends to have smaller RMSE due to the smaller estimated variance, relative to the extreme allocation, $n_A = 900$ and $n_B = 100$. Slight differences can be noted between the RMSE for the proportional allocation, $n_A = 600$ and $n_B = 400$ and the equal allocation, $n_A = 500$ and $n_B = 500$, due to the small difference in sample size allocations. A similar sample allocation effect applies under the GLR model in Table 2.6 as well.

As indicated in Figure 2.3, under simple sampling design from the GLR model population, where $n_A = 500$ and $n_B = 500$ and $\beta_a = 5$, $\beta_{ab} = 6$ and $\beta_b = 7$, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. This means that calibrating the 'stacked' samples directly by *z* or *x* in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ is not a satisfactory method for providing estimates from the dual frame sample. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$, as in Figure 2.3 and Table 2.5, resulted in reduced relative biases. Thus calibrating the 'stacked' samples directly by strata *J* and *z* or *x* in $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ is a more satisfactory way to combine the two samples under the complete response assumption.

Under the GLR model, in Table 2.5, the domain means do not have any effect on the relative biases in $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.xz}$. As in Figure 2.3, the higher the correlation between *y* and *x* the lower the relative biases achieved in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Generally, the RMSE, in Figure 2.3 and Table 2.6, show the same patterns as the relative biases.

Figure 2.3: Simulation RB (%) and RMSE (%) for FWE and JCE estimators estimated from the GLR model population under simple sampling design, equal allocation, and small domain mean differences

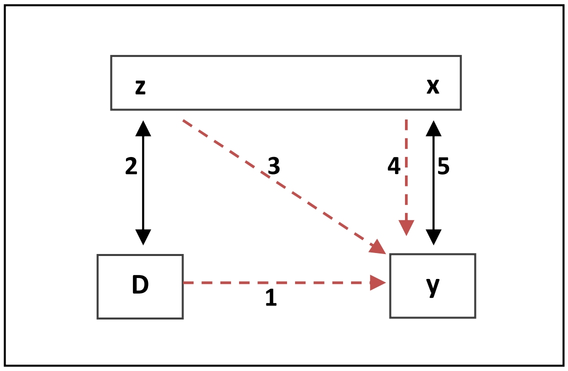


2.9 Discussion and Conclusion

In dual frame designs, three groups of variables may contribute to the estimation problem. The first group includes the study variables, $\mathbf{y}$. The second group includes the auxiliary variables, which might be associated with the study variable, such as $\mathbf{x}$, or might not, such as $\mathbf{z}$. The design domains, D, are the third group. Thinking about the relationship between these groups or variables can guide understanding about more satisfactory dual frame estimation approaches. Figure 2.4 shows the relationship between these different variables as studied in the simulation, where the bidirectional arrows indicate the association between two variables and the dashed arrows indicate different estimation scenarios.

Regardless of the relation between y and D, when accurate information about the design domains, D , is available, adding this information to the JCE auxiliary variable vector results in an unbiased estimate of the population total of y, (see arrow 1 in Figure 2.4). This is due to the fact that adding D to the auxiliary variable vector results in an estimator which is identical to the standard FWE dual frame estimator with $\theta = 0.5$. When a strong relationship exists between $\mathbf{z}$ and D (arrow 2, Figure 2.4), adding $\mathbf{z}$ to the JCE auxiliary variable vector results in reduced-biased estimates of Y (arrow 3, Figure 2.4). When a strong association exists between $\mathbf{x}$ and y (arrow 5, Figure 2.4), adding $\mathbf{x}$ to the JCE auxiliary variable vector results in an almost unbiased estimate of Y (arrow 4, Figure 2.4). Moreover, if adding $\mathbf{x}$ to the auxiliary variable vector results in a calibration model that closely matches the population model, this results in unbiased estimates of Y.

Figure 2.4: The relations between the study variable (*y*), auxiliary variables (x, z) and design domains (*D*) as guidance for dual frame estimation



In summary, the JCE was proposed here as a new model-assisted design-based dual frame estimator that can achieve parallel efficiency to that of the standard dual frame estimators. JCE has achieved a level of bias and MSE that is comparable to the standard estimator, FWE, under the assumption of complete response. JCE for point estimates is also easier to use in practice. Moreover, applying JCE does not necessarily require any information about the design domain membership, information required for standard dual frame estimators.

Generally, the performance of JCE depends on the agreement between the population model and the working model in the calibration, and to a lesser degree, on the association between the auxiliary variable and the study variable. Under the complete response assumption, when the auxiliary vector or the implicit calibration model more closely matches the population model, JCE yields almost unbiased dual frame estimates. When the models do not agree, JCE has a higher level of bias than the standard FWE estimator. Thus, the extent of the association between the study variable *y* and the auxiliary variable *x* is an important determinant factor of the JCE performance.

JCE ought to be preferred to the standard dual frame estimators, since it only depends on calibrating the pooled datasets to available auxiliary variables, a step already performed in standard dual frame estimation. In JCE, practitioners only need to apply the raking or poststratification step using available auxiliary variables, which are most likely related to the study variables. In addition, unlike the optimal dual frame estimators, JCE yields only one weighting variable to be used with the study variables. JCE can be easily extended to the multiple frame case; extending standard dual frame estimators to the multiple frame design is not readily done.

Finally, the JCE dual frame estimator has five of the desirable properties discussed in Chapter 1. It is unbiased or approximately so, internally consistent, efficient, applicable for point estimates with standard survey software and applicable to multiple (more than two) frame surveys. The other desirable properties will be explored in the subsequent chapters.

Table 2.3: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under simple sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	0	-0.01	0.01	-0.01	-0.03	-0.01	-0.03
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	0.05	0.02	0.03	-0.01	-0.03	-0.01	-0.03
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	0.05	0.04	0.07	0.03	0.01	0.02	0
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	0.03	0.02	0.06	0.06	0.08	0.06	0.08
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	0.07	0.04	0.05	0.07	0.06	0.07	0.06
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	0.01	0	0.08	0	0.05	0	0.05
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.03	-0.04	-0.07	-0.05	-0.06	-0.05	-0.06
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.02	-0.05	-0.06	-0.07	-0.05	-0.07	-0.05
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.01	-0.02	-0.05	-0.03	-0.03	-0.04	-0.04
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.04	-0.05	-0.08	-0.02	-0.04	-0.02	-0.04
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.07	-0.1	-0.09	-0.04	-0.02	-0.04	-0.02
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.13	-0.13	-0.11	-0.07	-0.03	-0.07	-0.03
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	0	-0.01	0	0.06	0.08	0.06	0.08
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	0.05	0.02	0.06	0.04	0.07	0.04	0.07
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.02	-0.02	0	0.02	0.06	0.02	0.07
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	0.1	0.09	0.09	0.07	0.09	0.07	0.09
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	0.15	0.12	0.11	0.07	0.07	0.07	0.07
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	0.12	0.12	0.07	0.05	0.03	0.05	0.03
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	0.01	0	0.07	-0.05	-0.01	-0.05	-0.02
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	0.05	0.02	0.1	-0.04	0	-0.04	0
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	-0.05	-0.06	0.02	-0.02	0.02	-0.03	0.02
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	0.04	0.03	0.07	-0.02	0.02	-0.02	0.02
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	-0.08	-0.11	-0.08	-0.05	-0.02	-0.05	-0.02
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	0.06	0.06	0.09	0.06	0.1	0.06	0.1
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	-0.09	-0.1	-0.12	-0.04	-0.09	-0.04	-0.09
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	0.03	0	0	0.01	-0.02	0.01	-0.02
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	0.05	0.05	0.05	0	-0.05	0	-0.05

Table 2.4: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under simple sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	2.27	1.93	1.8	1.74	1.63	1.74	1.64
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	2.21	1.93	1.82	1.74	1.63	1.74	1.64
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	2.88	2.51	2.56	2.29	2.33	2.29	2.33
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	2.3	1.96	1.79	1.81	1.65	1.81	1.65
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	2.25	1.96	1.85	1.78	1.67	1.78	1.67
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	2.86	2.49	2.55	2.28	2.33	2.28	2.33
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	2.24	1.95	1.82	1.76	1.65	1.76	1.65
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	2.14	1.88	1.77	1.72	1.64	1.72	1.64
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	2.78	2.42	2.49	2.18	2.22	2.18	2.23
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	2.29	2	1.84	1.62	1.49	1.62	1.49
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	2.16	1.9	1.79	1.54	1.44	1.54	1.44
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	2.7	2.35	2.4	1.88	1.91	1.88	1.92
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	2.27	1.94	1.79	1.58	1.44	1.58	1.44
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	2.19	1.88	1.79	1.55	1.46	1.55	1.46
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	2.71	2.34	2.4	1.89	1.93	1.88	1.93
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	2.18	1.87	1.74	1.51	1.41	1.52	1.41
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	2.13	1.79	1.7	1.45	1.38	1.44	1.38
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	2.81	2.44	2.51	1.92	1.98	1.92	1.99
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	2.3	2.01	1.87	1.21	1.14	1.22	1.14
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	2.21	1.91	1.78	1.12	1.07	1.12	1.07
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	2.93	2.58	2.63	1.54	1.58	1.54	1.58
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	2.33	2.04	1.88	1.21	1.1	1.21	1.1
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	2.26	1.98	1.86	1.18	1.1	1.18	1.1
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	2.85	2.51	2.59	1.54	1.62	1.54	1.62
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	2.32	2.06	1.86	1.21	1.11	1.21	1.11
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	2.23	1.99	1.87	1.17	1.11	1.17	1.11
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	2.85	2.51	2.59	1.49	1.57	1.49	1.57

Table 2.5: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under simple sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	0.02	0.02	5.76	-0.03	3.8	-0.02	3.8	-0.05	-0.08	-0.08
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	0.01	0	5.77	-0.08	3.77	-0.07	3.77	-0.05	-0.11	-0.11
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	-0.01	0	5.74	-0.04	3.81	-0.04	3.8	-0.1	-0.11	-0.12
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	0.13	0.13	5.82	0.14	3.82	0.14	3.82	0.06	0.07	0.07
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	0.12	0.11	5.79	0.11	3.78	0.12	3.78	0.04	0.05	0.04
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	0.08	0.1	5.8	0.05	3.76	0.06	3.75	0.01	-0.02	-0.02
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	0.07	0.07	5.73	0.08	3.82	0.09	3.81	0.02	0.04	0.03
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	0.14	0.13	5.79	0.12	3.88	0.13	3.88	0.1	0.09	0.1
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	0.16	0.19	5.86	0.2	3.95	0.21	3.94	0.11	0.15	0.14
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	0.07	0.07	6.06	0.07	3.37	0.07	3.36	0.03	0.07	0.06
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	0.08	0.07	6.05	0.04	3.34	0.05	3.35	0.04	0.05	0.05
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.03	-0.01	5.97	0.03	3.36	0.04	3.35	-0.06	0.02	0.03
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	0.05	0.05	6.11	0.11	3.4	0.11	3.4	0.07	0.09	0.09
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.1	-0.11	5.93	0	3.31	0.01	3.31	-0.1	-0.01	-0.01
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.11	-0.09	5.92	-0.08	3.23	-0.06	3.24	-0.15	-0.13	-0.13
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	-0.12	-0.11	5.83	-0.04	3.24	-0.04	3.25	-0.19	-0.09	-0.08
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	0.01	-0.02	5.94	0.02	3.32	0.02	3.32	-0.07	-0.01	-0.01
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	0.04	0.06	5.98	0.07	3.35	0.07	3.35	-0.04	0.02	0.02
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	-0.13	-0.12	5.95	0.01	2.47	0.02	2.47	-0.19	-0.04	-0.04
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	-0.07	-0.09	5.97	0.01	2.47	0.01	2.47	-0.14	-0.04	-0.04
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	-0.09	-0.08	5.96	0.02	2.5	0.02	2.5	-0.19	-0.04	-0.04
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	0.02	0.03	6.15	-0.02	2.47	-0.01	2.47	-0.07	-0.03	-0.03
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	0.08	0.07	6.22	0.02	2.5	0.02	2.5	-0.01	0.01	0.01
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	0.13	0.16	6.34	0.08	2.57	0.08	2.57	0.06	0.06	0.07
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	0.07	0.08	6.13	0	2.47	0.01	2.47	-0.03	-0.04	-0.04
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	0.02	0	6.06	-0.03	2.44	-0.02	2.44	-0.07	-0.06	-0.06
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	0.09	0.11	6.13	0.1	2.58	0.1	2.58	0.02	0.06	0.06

Table 2.6: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under simple sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	2.44	2.45	6.19	2.22	4.3	2.22	4.31	2.33	2.15	2.15
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	2.26	2.27	6.16	2.1	4.26	2.1	4.26	2.15	2.01	2.02
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	2.86	2.96	6.5	2.67	4.69	2.67	4.68	2.89	2.65	2.65
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	2.47	2.47	6.23	2.17	4.28	2.17	4.28	2.33	2.09	2.09
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	2.32	2.3	6.17	2.11	4.25	2.12	4.25	2.19	2.03	2.04
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	3	3.09	6.58	2.73	4.67	2.73	4.66	3	2.71	2.71
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	2.38	2.45	6.17	2.2	4.33	2.2	4.33	2.3	2.11	2.11
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	2.26	2.31	6.19	2.09	4.37	2.09	4.37	2.19	2.03	2.03
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	2.95	3.14	6.67	2.73	4.81	2.72	4.81	2.99	2.68	2.68
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	2.29	2.32	6.42	1.81	3.77	1.8	3.76	2.13	1.73	1.72
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	2.28	2.25	6.4	1.76	3.73	1.76	3.73	2.12	1.7	1.69
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	2.83	2.93	6.68	2.29	4.12	2.29	4.12	2.9	2.32	2.33
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	2.35	2.42	6.49	1.92	3.82	1.92	3.82	2.24	1.83	1.83
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	2.28	2.33	6.32	1.81	3.73	1.81	3.73	2.17	1.73	1.73
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	2.83	2.96	6.63	2.25	3.97	2.25	3.97	2.88	2.26	2.27
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	2.34	2.35	6.21	1.83	3.66	1.83	3.67	2.21	1.76	1.76
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	2.25	2.3	6.32	1.79	3.72	1.79	3.72	2.16	1.7	1.7
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	2.87	3	6.72	2.28	4.08	2.28	4.07	2.93	2.29	2.29
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	2.33	2.38	6.34	1.44	2.8	1.44	2.8	2.23	1.37	1.37
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	2.22	2.24	6.33	1.39	2.8	1.4	2.8	2.1	1.33	1.33
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	2.94	3.08	6.73	1.81	3.12	1.81	3.12	3	1.8	1.8
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	2.43	2.53	6.57	1.47	2.81	1.47	2.82	2.35	1.39	1.39
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	2.26	2.36	6.6	1.4	2.84	1.4	2.84	2.19	1.34	1.34
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	2.88	3.04	7.05	1.75	3.14	1.75	3.14	2.95	1.75	1.76
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	2.42	2.46	6.53	1.39	2.79	1.39	2.79	2.27	1.33	1.33
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	2.3	2.35	6.44	1.37	2.76	1.37	2.76	2.18	1.3	1.31
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	2.97	3.14	6.9	1.73	3.15	1.73	3.14	2.98	1.73	1.73

Appendix 2.1: Proof of Proposition 2.1

Where the calibration estimator in (2.38) is equivalent to the GREG estimator in (2.34), the JCE can be written as

$$\hat{Y}_{JCE} = \sum_U \hat{y}_k + \sum_{s_A} d_k \left(y_k - \hat{y}_k\right) + \sum_{s_B} d_k \left(y_k - \hat{y}_k\right) \tag{2.49}$$

$$= \sum_U \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_A} d_k y_k - \sum_{s_A} d_k \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_B} d_k y_k - \sum_{s_B} d_k \mathbf{x}'_k \hat{B}_s^{A,B} \tag{2.50}$$

$$\hat{Y}_{JCE} - Y = \sum_U \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_A} d_k y_k - \sum_{s_A} d_k \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_B} d_k y_k - \sum_{s_B} d_k \mathbf{x}'_k \hat{B}_s^{A,B} - \sum_U y_k \tag{2.51}$$

$$\begin{aligned}\hat{Y}_{JCE} - Y &= \sum_U \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_A} d_k y_k - \sum_{s_A} d_k \mathbf{x}'_k \hat{B}_s^{A,B} + \sum_{s_B} d_k y_k - \sum_{s_B} d_k \mathbf{x}'_k \hat{B}_s^{A,B} - \sum_U y_k \\ &-\sum_U \mathbf{x}_k \mathrm{B}_U - \sum_{s_A} d_k \mathbf{x}'_k \mathrm{B}_U - \sum_{s_B} d_k \mathbf{x}'_k \mathrm{B}_U + \sum_U \mathbf{x}'_k \mathrm{B}_U + \sum_{s_A} d_k \mathbf{x}'_k \mathrm{B}_U + \sum_{s_B} d_k \mathbf{x}'_k \mathrm{B}_U\end{aligned} \tag{2.52}$$

where $e_k = y_k - \mathbf{x}'_k \mathrm{B}_U$ and $\mathrm{B}_U = \left(\sum_U \mathbf{x}_k y_k\right)\left(\sum_U \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$

$$\hat{Y}_{JCE} - Y = A + C \tag{2.53}$$

Where

$$A = \sum_{s_A} d_k e_k + \sum_{s_B} d_k e_k - \sum_U e_k$$

$$C = \left(\sum_U \mathbf{x}'_k - \sum_{s_A} d_k \mathbf{x}'_k - \sum_{s_B} d_k \mathbf{x}'_k\right)\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)$$

$$E\left(\hat{Y}_{JCE} - Y\right) = E\left(A\right) + E\left(C\right) \tag{2.54}$$

$$E\left(A\right) = \sum_{U_A} e_k + \sum_{U_B} e_k - \sum_U e_k = \sum_{U_{ab}} e_k \tag{2.55}$$

$$E\left(C\right) = E\left(\sum_U \mathbf{x}'_k - \sum_{s_A} d_k \mathbf{x}'_k - \sum_{s_B} d_k \mathbf{x}'_k\right).E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) \tag{2.56}$$

$$= -\sum_{U_{ab}} \mathbf{x}_k' . E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) \tag{2.57}$$

By Taylor Linearization, the estimator $\hat{B}_s^{A,B}$ can be defined as

$$\begin{aligned} \hat{B}_s^{A,B} &= \mathrm{B}_U^{A,B} + \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{s'} d_k \mathbf{x}_k y_k - \sum_{U'} \mathbf{x}_k y_k\right) \\ &\quad - \sum_{U'} \mathbf{x}_k y_k \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-2} \left(\sum_{s'} d_k \mathbf{x}_k \mathbf{x}_k' - \sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right) \end{aligned} \tag{2.58}$$

where

$$\sum_{s'} d_k \mathbf{x}_k y_k = \sum_{s_A} d_k \mathbf{x}_k y_k + \sum_{s_B} d_k \mathbf{x}_k y_k$$

$$\sum_{s'} d_k \mathbf{x}_k \mathbf{x}_k' = \sum_{s_A} d_k \mathbf{x}_k \mathbf{x}_k' + \sum_{s_B} d_k \mathbf{x}_k \mathbf{x}_k'$$

$$\sum_{U'} \mathbf{x}_k y_k = \sum_{U_A} \mathbf{x}_k y_k + \sum_{U_B} \mathbf{x}_k y_k = \sum_{U} \mathbf{x}_k y_k + \sum_{U_{ab}} \mathbf{x}_k y_k$$

$$\sum_{U'} \mathbf{x}_k \mathbf{x}_k' = \sum_{U_A} \mathbf{x}_k \mathbf{x}_k' + \sum_{U_B} \mathbf{x}_k \mathbf{x}_k' = \sum_{U} \mathbf{x}_k \mathbf{x}_k' + \sum_{U_{ab}} \mathbf{x}_k \mathbf{x}_k'$$

$$\hat{B}_s^{A,B} = \left(\sum_{s'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{s'} d_k \mathbf{x}_k y_k\right)$$

$$\mathrm{B}_U^{A,B} = \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{U'} \mathbf{x}_k y_k\right)$$

$$\begin{aligned} \hat{B}_s^{A,B} &= \mathrm{B}_U + \mathrm{B}_U^{A,B} - \mathrm{B}_U + \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{s'} d_k \mathbf{x}_k y_k - \sum_{U'} \mathbf{x}_k y_k\right) \\ &\quad - \sum_{U'} \mathbf{x}_k y_k \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-2} \left(\sum_{s'} d_k \mathbf{x}_k \mathbf{x}_k' - \sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right) \end{aligned} \tag{2.59}$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_U + \mathrm{B}_U^{A,B} - \mathrm{B}_U \tag{2.60}$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_\mathrm{U} + \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{U'} \mathbf{x}_k y_k\right) - \left(\sum_{U} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{U} \mathbf{x}_k y_k\right) \tag{2.61}$$

$$E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) = \left(\sum_{U'} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{U'} \mathbf{x}_k y_k\right) - \left(\sum_{U} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{U} \mathbf{x}_k y_k\right) \tag{2.62}$$

$$\therefore E\left(C\right) = -\sum\nolimits_{U_{ab}} \mathbf{x}_k' \left(\left(\sum\nolimits_{U'} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum\nolimits_{U'} \mathbf{x}_k y_k \right) - \left(\sum\nolimits_{U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum\nolimits_{U} \mathbf{x}_k y_k \right) \right) \tag{2.63}$$

Consequently, under dual frame design

$$E\left(\hat{Y}_{JCE} - Y\right) = \sum\nolimits_{U_{ab}} e_k - \sum\nolimits_{U_{ab}} \mathbf{x}_k' \left(\left(\sum\nolimits_{U'} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum\nolimits_{U'} \mathbf{x}_k y_k \right) - \left(\sum\nolimits_{U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum\nolimits_{U} \mathbf{x}_k y_k \right) \right) \tag{2.64}$$

$$= \sum\nolimits_{U_{ab}} \left(y_k - \mathbf{x}_k' \left(\sum\nolimits_{U'} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum\nolimits_{U'} \mathbf{x}_k y_k \right) \right) \tag{2.65}$$

$$= \sum\nolimits_{U_{ab}} \left(y_k - \mathbf{x}_k' \mathrm{B}_U^{A,B} \right) \tag{2.66}$$

$$\therefore B\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{U_{ab}} e_k^{A,B} \tag{2.67}$$

where $e_k^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_U^{A,B} \right)$

Appendix 2.2: Results for Complex Sampling Designs

Table 2.7: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under complex sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	0.13	0.1	0.13	0.13	0.13	0.13	0.13
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	0.11	0.06	0.05	0.05	-0.01	0.05	-0.01
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	0.06	0.07	0.08	0.02	-0.02	0.01	-0.02
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	0.01	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	0.06	0	0.02	0.01	0	0.01	0
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	0.01	0.02	0.08	-0.04	0.01	-0.04	0.01
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.02	-0.06	-0.12	-0.05	-0.1	-0.06	-0.1
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.01	-0.07	-0.1	-0.06	-0.06	-0.07	-0.07
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	0.02	0.02	-0.03	-0.01	-0.03	-0.02	-0.04
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	0.08	0.05	0.01	0.04	0.04	0.04	0.04
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	0.11	0.06	0.06	0.05	0.09	0.05	0.09
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.14	-0.13	-0.09	-0.13	-0.06	-0.13	-0.06
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.16	-0.2	-0.18	-0.16	-0.11	-0.16	-0.11
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	0.05	0.01	0.03	-0.04	0	-0.04	0
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.02	-0.01	0.02	-0.01	0.05	0	0.05
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	0.1	0.05	0.04	0.01	0.03	0.01	0.03
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	0.09	0.04	0.04	0.03	0.04	0.03	0.04
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	0.03	0.04	0.01	-0.03	-0.03	-0.03	-0.03
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	0.04	0	0.1	0	0.03	0	0.03
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	0.03	-0.02	0.07	0.02	0.05	0.02	0.05
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	-0.09	-0.08	0.01	-0.03	0.02	-0.03	0.02
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	0.1	0.05	0.08	-0.05	-0.03	-0.05	-0.03
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	0.05	0	-0.01	-0.02	0.01	-0.02	0.01
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	0.04	0.05	0.09	0.07	0.11	0.07	0.11
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	0.01	-0.03	-0.03	-0.07	-0.1	-0.08	-0.1
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	0.01	-0.04	0.03	0.02	0.01	0.02	0.01
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	0.01	0.02	0.05	-0.05	-0.1	-0.05	-0.1

Table 2.8: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under complex sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	2.3	2.06	1.96	1.91	1.81	1.91	1.81
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	2.24	2.09	1.99	1.91	1.81	1.92	1.81
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	2.78	2.44	2.51	2.24	2.29	2.23	2.29
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	2.29	2.11	1.94	1.92	1.77	1.92	1.77
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	2.31	2.13	2.02	1.93	1.83	1.93	1.83
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	2.83	2.5	2.56	2.27	2.33	2.26	2.33
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	2.22	2.04	1.93	1.91	1.81	1.91	1.82
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	2.2	2.03	1.96	1.86	1.79	1.86	1.79
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	2.79	2.53	2.61	2.28	2.34	2.28	2.34
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	2.28	2.09	1.96	1.68	1.6	1.68	1.6
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	2.19	1.99	1.9	1.58	1.5	1.58	1.5
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	2.73	2.45	2.55	1.98	2.07	1.98	2.07
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	2.27	2.11	1.97	1.69	1.57	1.69	1.57
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	2.18	2.02	1.93	1.66	1.57	1.66	1.57
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	2.75	2.51	2.6	2.04	2.11	2.04	2.11
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	2.2	2.04	1.88	1.65	1.52	1.65	1.53
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	2.13	1.89	1.84	1.49	1.44	1.49	1.44
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	2.71	2.44	2.54	1.93	1.99	1.93	1.99
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	2.26	2.11	2.02	1.3	1.25	1.3	1.25
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	2.28	2.09	1.95	1.2	1.16	1.2	1.16
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	2.9	2.61	2.66	1.56	1.62	1.55	1.61
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	2.36	2.17	2.07	1.34	1.3	1.34	1.3
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	2.27	2.08	2	1.25	1.19	1.25	1.19
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	2.86	2.62	2.71	1.56	1.62	1.56	1.62
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	2.46	2.3	2.15	1.35	1.28	1.35	1.28
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	2.31	2.14	2.03	1.28	1.22	1.28	1.23
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	2.87	2.55	2.62	1.53	1.58	1.53	1.59

Table 2.9: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under complex sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	0.17	0.14	5.9	0.11	3.93	0.11	3.92	0.07	0.06	0.06
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	0.05	0.01	5.79	-0.02	3.82	-0.01	3.82	-0.04	-0.05	-0.05
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	-0.02	0	5.75	0.01	3.86	0.01	3.85	-0.11	-0.07	-0.07
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	0.03	-0.01	5.73	0.07	3.8	0.07	3.8	-0.06	0.01	0.01
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	0.11	0.07	5.78	0.07	3.77	0.07	3.77	0.02	0.02	0.02
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	-0.08	-0.05	5.66	-0.09	3.64	-0.08	3.64	-0.13	-0.15	-0.15
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.02	-0.05	5.6	-0.05	3.67	-0.05	3.67	-0.11	-0.11	-0.11
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	0.01	-0.02	5.69	-0.06	3.73	-0.05	3.74	-0.05	-0.09	-0.08
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	-0.06	-0.03	5.68	0.04	3.81	0.05	3.81	-0.08	0.01	0.01
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	0.15	0.12	6.1	0.04	3.33	0.04	3.33	0.11	0.05	0.05
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	0.1	0.06	6.06	0.05	3.36	0.05	3.36	0.06	0.06	0.06
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	-0.04	-0.01	5.97	0.05	3.37	0.06	3.37	-0.04	0.04	0.05
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	0.14	0.11	6.13	0.01	3.32	0.02	3.32	0.12	0.03	0.03
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.11	-0.15	5.89	-0.02	3.29	-0.01	3.29	-0.11	-0.01	-0.01
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	-0.14	-0.11	5.94	-0.08	3.26	-0.07	3.26	-0.13	-0.11	-0.1
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	0.21	0.18	6.07	0.07	3.35	0.07	3.35	0.1	0.05	0.04
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	0.15	0.11	6.01	0.02	3.3	0.01	3.3	0.04	0.01	0.01
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	0	0.03	5.99	0	3.33	0	3.32	-0.04	-0.02	-0.02
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	0.13	0.09	6.13	0.02	2.49	0.02	2.49	0.03	-0.01	-0.01
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	0.09	0.05	6.09	-0.01	2.46	0	2.46	0.02	-0.03	-0.03
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	0.01	0.03	6.03	0.03	2.51	0.02	2.5	-0.06	-0.01	-0.01
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	-0.1	-0.13	6.02	-0.1	2.4	-0.1	2.4	-0.18	-0.09	-0.09
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	-0.03	-0.06	6.11	-0.07	2.42	-0.07	2.42	-0.1	-0.07	-0.07
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	0.13	0.16	6.36	0.02	2.54	0.03	2.55	0.09	0.03	0.04
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	0.05	0.01	6.04	0.03	2.51	0.03	2.5	-0.09	0	-0.01
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	0.05	0.01	6.06	0.04	2.51	0.04	2.51	-0.08	-0.01	-0.01
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	0.15	0.18	6.19	0.12	2.6	0.12	2.6	0.08	0.09	0.08

Table 2.10: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under complex sampling design.

Sample size	Domain means	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.40	2.62	2.51	6.36	2.31	4.5	2.32	4.51	2.4	2.22	2.22
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.40	2.49	2.42	6.23	2.23	4.37	2.23	4.38	2.3	2.15	2.15
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.40	3.02	3.08	6.57	2.76	4.8	2.76	4.79	2.99	2.74	2.75
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.40	2.59	2.53	6.21	2.27	4.36	2.27	4.36	2.38	2.15	2.15
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.40	2.51	2.47	6.26	2.23	4.36	2.22	4.36	2.37	2.17	2.17
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.40	3.08	3.1	6.48	2.79	4.63	2.79	4.63	3.07	2.78	2.79
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.40	2.68	2.64	6.09	2.3	4.24	2.31	4.24	2.46	2.18	2.19
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.40	2.56	2.52	6.17	2.29	4.33	2.29	4.33	2.38	2.18	2.19
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.40	2.97	3.08	6.51	2.75	4.73	2.74	4.73	2.98	2.7	2.7
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.60	2.41	2.35	6.5	1.92	3.8	1.92	3.8	2.22	1.83	1.83
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.60	2.36	2.31	6.44	1.84	3.8	1.84	3.79	2.18	1.76	1.76
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.60	2.87	2.9	6.66	2.29	4.12	2.29	4.12	2.87	2.32	2.32
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.60	2.42	2.4	6.52	1.94	3.78	1.94	3.79	2.23	1.82	1.82
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.60	2.4	2.4	6.33	1.92	3.77	1.92	3.77	2.29	1.84	1.84
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.60	3.01	3.03	6.69	2.39	4.07	2.39	4.07	3	2.41	2.41
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.60	2.55	2.48	6.52	1.96	3.84	1.96	3.84	2.34	1.88	1.88
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.60	2.39	2.38	6.42	1.87	3.76	1.87	3.75	2.24	1.79	1.79
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.60	2.93	3.01	6.73	2.24	4.05	2.24	4.04	2.91	2.25	2.24
n1=500, n2=500	β_d =(5,6,7)	ρ_{xy} = 0.80	2.55	2.45	6.55	1.48	2.86	1.48	2.86	2.33	1.41	1.41
n1=600, n2=400	β_d =(5,6,7)	ρ_{xy} = 0.80	2.39	2.36	6.49	1.47	2.83	1.47	2.83	2.22	1.38	1.38
n1=900, n2=100	β_d =(5,6,7)	ρ_{xy} = 0.80	3.01	3.12	6.82	1.83	3.14	1.83	3.13	3.01	1.82	1.82
n1=500, n2=500	β_d =(5,5,10)	ρ_{xy} = 0.80	2.6	2.53	6.49	1.47	2.77	1.47	2.77	2.38	1.41	1.41
n1=600, n2=400	β_d =(5,5,10)	ρ_{xy} = 0.80	2.4	2.34	6.51	1.45	2.8	1.45	2.8	2.19	1.39	1.39
n1=900, n2=100	β_d =(5,5,10)	ρ_{xy} = 0.80	3.11	3.2	7.15	1.86	3.17	1.86	3.18	3.07	1.84	1.84
n1=500, n2=500	β_d =(5,10,15)	ρ_{xy} = 0.80	2.52	2.42	6.47	1.42	2.85	1.42	2.85	2.3	1.35	1.35
n1=600, n2=400	β_d =(5,10,15)	ρ_{xy} = 0.80	2.43	2.4	6.46	1.42	2.84	1.42	2.84	2.25	1.35	1.35
n1=900, n2=100	β_d =(5,10,15)	ρ_{xy} = 0.80	3.04	3.22	6.98	1.81	3.2	1.81	3.2	3.07	1.79	1.79

Chapter 3

Joint Calibration Estimator in the Presence of Nonresponse Errors

3.1 Introduction

Chapter 2 provided an overview of the calibration approach and introduced the Joint Calibration Estimator (JCE) for dual frame estimation. The properties of the JCE were explored under the full response assumption; the simulation studies in Chapter 2 assumed the absence of the non-sampling errors. In this chapter, our concern is with dual frame surveys affected by unit nonresponse. That is, the values of the study variable *y* are not observed for all elements in the full samples s_A and s_B; *y* values are observed only for the elements in response sets r_A and r_B of the full samples s_A and s_B, respectively. In this chapter, we assume that *y* is affected by unit nonresponse only, and from now on, we will use 'nonresponse' instead of 'unit nonresponse'.

In this chapter, the JCE performance is explored in the presence of nonresponse. The JCE is introduced in this chapter as both a dual frame estimator and an approach for nonresponse adjustment. The nonresponse problem is discussed in Section 3.2. The JCE in the presence of nonresponse is presented in Sections 3.3 and 3.4. The bias for JCE in the presence of nonresponse is derived in Section 3.5. The performance of JCE in comparison with standard FWE dual frame estimator in the presence of nonresponse is explored by a simulation study described in Section 3.6. The simulation results and findings are presented and discussed in Sections 3.7 and 3.8.

3.2 Nonresponse in Dual Frame Design

As discussed in Chapter 1, the standard dual frame estimators assume that for each sample, the estimators of population totals are approximately unbiased for the corresponding domain total. This means that $\hat{Y}_a$ and $\hat{Y}_b$ are unbiased estimates for Y_a and Y_b, respectively, and both $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ are unbiased for Y_{ab}. Based on this assumption, either (1.3) or (1.4) achieves unbiased dual frame estimates, which is the first desirable property for dual frame estimators, as discussed in Chapter 1. Any violations of the unbiased domain estimates assumption results in biased dual frame estimates, which contradicts three of the eight desirable properties for dual frame estimators. These three properties are: 1) unbiasedness, 2) efficiency and 3) robustness. Biased domain estimates can result from several sources of non-sampling error, including nonresponse, non-coverage and misclassification (measurement) errors. In this chapter, we focus on the nonresponse error while in Chapter 4 we will explore misclassification error as a form of measurement error.

In dual frame designs, the nonresponse problem tends to be more complicated than for single frame designs since the samples from the different frames can have different nonresponse properties. For example, Brick, Dipko, Presser, Tucker, and Yuan (2006) showed that in dual frame telephone surveys, samples from the Random-Digit-Dialing (RDD) landline frames and cell phone frames can suffer from differential nonresponse due to noncontact resulting from differential accessibility. Differential accessibility occurs as a result of individuals' telephone usage preferences or patterns. Some studies have indicated that this kind of nonresponse bias could be more severe than the non-coverage error resulting from non-coverage of the non-landline households in the RDD landline telephone surveys (Brick et al., 2006; Kennedy, 2007).

In order to adjust for nonresponse in dual frame designs, the estimation problem will not be as straightforward as simply applying the design weights and combining the samples. An adjustment step is necessary either before or after the combining step. Consequently, the dual frame estimator might have a different form other than (1.3) or (1.4). For example, Brick et al. (2011) proposed a post-stratified estimator as a method to treat the differential accessibility nonresponse problem in dual frame telephone surveys. A composite factor $\hat{\theta}_0 = RR_l(RR_c - RR_{c.lm})/(RR_c RR_{l.lm} - RR_l RR_{c.lm})$ was proposed to reduce the differential accessibility nonresponse bias. RR_l , $RR_{l.lm}$, RR_c and $RR_{c.lm}$ are the response rates among the landline sample dual users, the landline sample landline-mainly users, the cell sample dual users, and the cell phone sample landline-mainly users, respectively. Identifying the landline-mainly users (persons who predominantly use landline telephones although they have cell phones) requires collecting more data during the interview about the households' telephone usage patterns. Moreover, no exact information about $RR_{l.lm}$ and $RR_{c.lm}$ is available, and only estimated response rates, $\hat{RR}_{l.lm}$ and $\hat{RR}_{c.lm}$, can be used.

If frame A denotes the RDD landline frame and frame B denotes the cell phone frame, Brick and his colleagues (2011) identified the telephone service post-stratified dual frame estimator as

$$\hat{Y}_{ps} = \frac{N_a}{\hat{N}_a}\hat{Y}_a + \theta_0 \frac{N_{ab}}{\hat{N}_{ab}^A}\hat{Y}_{ab}^A + \left(1-\theta_0\right)\frac{N_{ab}}{\hat{N}_{ab}^B}\hat{Y}_{ab}^B + \frac{N_b}{\hat{N}_b}\hat{Y}_b \tag{3.1}$$

Even more complicated estimators that explicitly incorporated the landline mainly and cell mainly dual user domain estimates have been proposed by Brick et al. (2011). This means that the nonresponse problem not only affects the desirable properties of the dual frame estimators but also the weight adjustment prescribed in the standard dual frame estimators.

Under single frame designs, the *stochastic model* of survey response views the response set $r\left(r \subseteq s\right)$ as the result of two probabilistic selections. In the first, sample *s* is selected from population *U*, and in the second, a response set *r* is realized as a subsample from the sample *s*. Two approaches for adjusting the nonresponse that fall under the *stochastic model* label are the *two-phase approach* and the *calibration approach* (Deville & Särndal, 1992; Lundström & Särndal, 1999; Särndal & Lundström, 2005; Särndal, Swensson, & Wretman, 1992). In the *two-phase approach*, assuming that the conditional response distribution, $q\left(r \middle| s\right)$, is known, the first-order response probabilities $\Pr\left(k \in r \middle| s\right) = \phi_k$ are known and can be used to adjust for nonresponse. Using the adjusted design weights d_k/ϕ_k, the unbiased two-phase estimator of *y* total can be written as

$$\hat{Y} = \sum_r \left(d_k/\phi_k\right) y_k \tag{3.2}$$

In practice, the response probabilities ϕ_k are unknown and the estimated response probabilities $\hat{\phi}_k$ can be used instead to adjust the base weights d_k. To estimate the response probabilities, assumptions about the response mechanism are needed. Based on the presumed response mechanism, a response model is formulated to estimate the response probabilities ϕ_k. For example, Little (1986) and Ekhlom and Laaksonen (1991) estimated the response probabilities from logistic regression models. Since this requires modeling the response mechanism, the *two-phase approach* is a population based method to adjust for nonresponse, which means that it requires case-level information for the nonrespondents. On the other hand, Lundström and Särndal (1999) suggested the *calibration approach* as a simple and unified sample based method to reduce both the nonresponse bias and variance. It adjusts the respondents directly using the available auxiliary variable totals. In addition to the lack of any needed pre-modeling steps in the *calibration approach*, the auxiliary information is needed only for the responding elements $k \in r$. As a property of the *calibration approach*, the auxiliary variables should contribute to reducing the nonresponse bias and the variance of estimates, as well (Chang & Kott, 2008; Kott, 2006; Kott & Chang, 2010; Särndal & Lundström, 2005). In practice, post-stratification and weighting-class adjustment, which are special cases of calibration, are used extensively to adjust for nonresponse (Lohr, 1999). Although the choice of calibration variables does not require an explicit modeling step, implicit modeling would help in justifying the selection of the auxiliary variables to be used.

Under dual frame designs and the *stochastic model* of survey response, the same two approaches for adjusting for nonresponse can be identified, the *two-phase approach* and the *calibration approach*. In the *two-phase approach*, the response mechanism or model is required for each separate frame. This requires developing the mathematical formulation of the response model in each sample and selecting the explanatory variables for these models from the available auxiliary variables for that frame. Under the *calibration approach*, the dual frame samples can be calibrated separately, before combining the two samples, or jointly, after combining the two samples (Lohr, 2011). In the following section, the JCE is introduced as a dual frame estimator that waives the combination step and uses a single set of available auxiliary variable totals to adjust for nonresponse in the combined dual frame sample.

Under the calibration approach, the nonresponse mechanism is assumed to be a function of a set of covariates, the *model variables*. These covariates may or may not coincide with the calibration *benchmark variables* in the calibration equation (Chang & Kott, 2008; Kott, 2006; Kott & Chang, 2010). In this chapter, we will assume that both model and benchmark variables coincide and that these variables are available only for respondents, $k \in r$ where *r* is response set $r\left(r \subseteq s\right)$, and not available for $k \in U - r$.

3.3 The Joint Calibration in the Presence of Nonresponse

When nonresponse is present in a single frame design, let $\mathbf{x}_k = \left(x_{k1}, .., x_{kj}, .., x_{kJ}\right)'$ denote the auxiliary variables vector observed for the response set elements $k \in r$, and assuming the corresponding auxiliary population totals $\mathbf{X} = \left(\sum_U x_{k1}, .., \sum_U x_{kj}, .., \sum_U x_{kJ}\right)'$ are known, the joint calibration problem is to find final weights w_k, $k \in r$, that satisfy the calibration equation

$$\sum\nolimits_r w_k \mathbf{x}_k = \mathbf{X} \tag{3.3}$$

through minimizing the distance function

$$\sum\nolimits_r \left(w_k - d_k\right)^2 \Big/ 2d_k \tag{3.4}$$

The dual frame design yields response sets $r_A\left(r_A \subseteq s_A\right)$ and $r_B\left(r_B \subseteq s_B\right)$, which when combined yield $r=\left(r_A,r_B\right)$. Similarly, $r_a\left(r_a \subseteq s_a\right)$, $r_b\left(r_b \subseteq s_b\right)$, $r_{ab}^A\left(r_{ab}^A \subseteq s_{ab}^A\right)$ and $r_{ab}^B\left(r_{ab}^B \subseteq s_{ab}^B\right)$ denote dual frame domain response sets. Where the calibration equation in (3.3) can be written as $\sum_{r_A} w_k \mathbf{x}_k + \sum_{r_B} w_k \mathbf{x}_k = \mathbf{X}$, the distance function in (3.4) can be split into two components

$$\sum\nolimits_{r_A}\left(w_k-d_k\right)^2 \Big/ 2d_k + \sum\nolimits_{r_B}\left(w_k-d_k\right)^2 \Big/ 2d_k \tag{3.5}$$

Using a Lagrange multiplier to obtain a minimum distance measure $G\left(d_k,w_k\right)$ between d_k and w_k under the calibration constraints, the joint calibration weights are

$$w_k = \begin{cases} d_k v_k & k \in r_A \\ d_k v_k & k \in r_B \end{cases} \tag{3.6}$$

Where $v_k = \left(1+\lambda_r' \mathrm{x}_k\right)$ is the joint calibration factor in the presence of nonresponse and $\lambda_r' = \left(\sum_U \mathbf{x}_\mathrm{k} - \sum_r d_k \mathbf{x}_\mathrm{k}\right)' \left(\sum_r d_k \mathbf{x}_\mathrm{k} \mathbf{x}_\mathrm{k}'\right)^{-1}$, the resulting JCE estimator is

$$\hat{Y}_{JCE} = \sum\nolimits_r w_k y_k \tag{3.7}$$

Note that under the single frame design, the final calibration weights w_k are calibrated to the auxiliary information and may implicitly account for nonresponse. Särndal and Lündstrom (2005, 2010) classified the auxiliary information $\mathbf{x}_k$ into two kinds, $\mathbf{x}_k^*$ and $\mathbf{x}_k^{\mathrm{o}}$. The population auxiliary information, $\mathbf{x}_k^*$, which is known for every $k \in U$ and the sample auxiliary information, $\mathbf{x}_k^{\mathrm{o}}$, which is known only for $k \in s$. Therefore, if both kinds are included in the calibration auxiliary information $\mathbf{x}_k$, the calibration weights w_k imply that $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ and $\sum_r w_k \mathbf{x}_k^{\circ} = \sum_s d_k \mathbf{x}_k^{\circ}$, where $\sum_s d_k \mathbf{x}_k^{\circ}$ is unbiased estimate for $\sum_U \mathbf{x}_k^{\circ}$.

In the presence of nonresponse, the Joint Calibration approach is motivated by *Remark 6.3* in Särndal and Lundström (2005). Based on the remark and under a single frame design, when a perfect linear relationship exists in the population between the study variable y_k and the auxiliary vector $\mathbf{x}_k$,

$$y_k = \mathbf{x}_k' \beta, \quad \text{for every } k \in U \tag{3.8}$$

where β is a column vector of unknown constants, then the calibration estimator $\hat{Y}_w$ gives an exact estimate of the target total *Y* as

$$\hat{Y}_w = \sum_r w_k y_k = \left(\sum_r w_k \mathbf{x}_k\right)' \beta = \left(\sum_U \mathbf{x}_k\right)' \beta = \sum_U y_k = Y \tag{3.9}$$

Although the perfect linear relationship in (3.9) does not hold in practice, this result suggests that using a powerful auxiliary vector $\mathbf{x}_k$ which has a strong linear relationship with the study variable y_k results in a calibration estimator $\hat{Y}_w$ that will closely approximate the target population total *Y*. The same argument holds under dual frame designs where (3.9) can be written as

$$\hat{Y}_w = \sum_{r_A} w_k y_k + \sum_{r_B} w_k y_k = \left(\sum_U \mathbf{x}_k\right)' \beta = \sum_U y_k = Y \tag{3.10}$$

However, note that this property is conditional on the agreement between the calibration model and the population model as proven for the full response condition in Appendix 2.1 in Chapter 2.

3.4 Examples of Joint Calibration Estimators in the Presence of Nonresponse

The auxiliary variable vector characterizes the final JCE for dual frame estimation. For example, under a univariate auxiliary variable $\mathbf{x}_k = 1$ for all $k \in U$, which corresponds to the *common mean model*, where the overall population total is $\mathbf{X} = N$, the joint calibration factor is

$$v_k = N\left(\sum\nolimits_{r_A} d_k + \sum\nolimits_{r_B} d_k\right)^{-1} \tag{3.11}$$

By calibrating the combined datasets, $\sum\nolimits_{r_A} w_k \mathbf{x}_k + \sum\nolimits_{r_B} w_k \mathbf{x}_k = N$. Then $\hat{Y}_{JCE}$ can be written as

$$\hat{Y}_{JCE} = N\left(\sum\nolimits_{r_A} d_k y_k + \sum\nolimits_{r_B} d_k y_k\right)\left(\sum\nolimits_{r_A} d_k + \sum\nolimits_{r_B} d_k\right)^{-1} \tag{3.12}$$

Under the *Simple Random Sampling* (SRS) design, $\hat{Y}_{JCE}$ in (3.12) can be written as

$$\hat{Y}_{JCE} = N \frac{m_A f_B \bar{y}_r^A + m_B f_A \bar{y}_r^B}{m_A f_B + m_B f_A} \tag{3.13}$$

where the sampling fractions are $f_A = n_A / N_A$ and $f_B = n_B / N_B$ for samples A and B, respectively, and where $\bar{y}_r^A$ and $\bar{y}_r^B$ are the estimated means $\sum\nolimits_{r_A} y_k / m_A$ and $\sum\nolimits_{r_B} y_k / m_B$ from the respondents m_A and m_B for samples A and B, respectively. If the nonresponse is not completely at random across the two samples, calibration under the *common mean model* will not adjust for nonresponse bias. However, $\hat{Y}_{JCE}$ in (3.12) still finds use if there are no better auxiliary variables; it gives better estimates, with reduced bias, than the non-calibrated standard dual frame estimators.

Where $\mathbf{x}_k = x_k$, for all $k \in U$, which corresponds to *the ratio model*, $\mathbf{X} = X = \sum_U x_k$, the joint calibration factor can be written as

$$v_k = X\left(\sum_{r_A} d_k^A x_k + \sum_{r_B} d_k^B x_k\right)^{-1} \tag{3.14}$$

By calibrating the combined datasets, $\sum_{r_A} w_k \mathbf{x}_k + \sum_{r_B} w_k \mathbf{x}_k = X$. Then $\hat{Y}_{JCE}$ can be written as

$$\hat{Y}_{JCE} = X\left(\sum_{r_A} d_k y_k + \sum_{r_B} d_k y_k\right)\left(\sum_{r_A} d_k x_k + \sum_{r_B} d_k x_k\right)^{-1} \tag{3.15}$$

Under the *Simple Random Sampling* design, $\hat{Y}_{JCE}$ in (3.15) can be written as

$$\hat{Y}_{JCE} = N\bar{X}\,\frac{m_A f_B \bar{y}_r^A + m_B f_A \bar{y}_r^B}{m_A f_B \bar{x}_r^A + m_B f_A \bar{x}_r^B} \tag{3.16}$$

where $\bar{X} = \sum_U x_k / N$ and $\bar{x}_r^A$ and $\bar{x}_r^B$ are the estimated means $\sum_{r_A} x_k / m_A$ and $\sum_{r_B} x_k / m_B$ from the respondents m_A and m_B for samples A and B, respectively. If the nonresponse does not occur completely at random, joint calibration under the *ratio model* will not adjust for nonresponse bias.

Under the multivariate auxiliary variable $\mathbf{x}_k = (1, x_k)$ for all $k \in U$, which corresponds to the *simple regression model with intercept*, where x_k is the value for element *k* of a continuous variable *x*, and the population total vector is $\mathbf{X} = (N, X)$, $\hat{Y}_{JCE}$ can be written as

$$\hat{Y}_{JCE} = \hat{Y}_{HT}^{A} + \hat{Y}_{HT}^{B} + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{r_A} d_k x_k + \sum\nolimits_{r_B} d_k x_k\right)\right)\hat{B}_r \tag{3.17}$$

where $\hat{B}_r = \sum_r d_k \mathbf{x}_k y_k \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$. This estimator gives better protection against nonresponse bias than the *ratio model* estimator, since the regression estimator is nearly unbiased under equal response probability within groups (Särndal & Lundström, 2005).

Another interesting multivariate calibration estimator is the complete post-stratified estimator, which corresponds to the *group mean model*, where the calibration is on known post-stratified cell counts. When the sizes of the population groups N_p are known and the classification vector used to code membership in one of P mutually exclusive and exhaustive groups $\mathbf{x}_k = \gamma_k = \left(\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk}\right)'$ is used as the auxiliary variables vector, the calibrated estimator will be the well-known post-stratified estimator.

Under the dual frame design, the joint calibration factor takes the following form

$$v_k = N_p \left(\sum\nolimits_{r_p^A} d_k + \sum\nolimits_{r_p^B} d_k\right)^{-1} \tag{3.18}$$

where r_p^A denotes the sample cell $U_p \cap r_A$ and r_p^B denotes the sample cell $U_p \cap r_B$. In this case, $\hat{Y}_{JCE}$ can be written as

$$\hat{Y}_{JCE} = \sum\nolimits_{P} N_p \left(\sum\nolimits_{r_p^A} d_k y_k + \sum\nolimits_{r_p^B} d_k y_k\right)\left(\sum\nolimits_{r_p^A} d_k + \sum\nolimits_{r_p^B} d_k\right)^{-1} \tag{3.19}$$

Under the *Simple Random Sampling* design, $\hat{Y}_{JCE}$ in (3.19) can be written as

$$\hat{Y}_{JCE} = \sum\nolimits_{P} N_p \frac{m_p^A f_p^B \overline{y}_{p;r}^A + m_p^B f_p^A \overline{y}_{p;r}^B}{m_p^A f_p^B + m_p^B f_p^A} \tag{3.20}$$

where f_p^A and f_p^B are sampling fractions within group *p* for samples A and B, respectively, and where $\overline{y}_{p;r}^A$ and $\overline{y}_{p;r}^B$ are the estimated means $\sum\nolimits_{r_p^A} y_k \Big/ m_p^A$ and $\sum\nolimits_{r_p^B} y_k \Big/ m_p^B$ from the respondents m_p^A and m_p^B for group *p* in samples A and B, respectively.

In the *group mean model*, it is implicitly assumed that mean and variance are the same for all elements within the same group *p*. Similarly, where the group totals X_p are known and $\mathbf{x}_k = x_k \gamma_k = \left(x_{1k}\gamma_{1k}, \ldots, x_{pk}\gamma_{pk}, \ldots, x_{Pk}\gamma_{Pk}\right)'$ is used as the auxiliary variable vector, this corresponds to the *group ratio model*, where ratio mean and variance are shared by all elements within the same group *p*. Both the *group mean model* and *group ratio model* may be classified as *group models*. Since the groups in the *group models* can serve as strata, JCE would have better performance if this informative design has been included in the auxiliary variables totals. At the same time, if nonresponse does not occur at random within every group, calibration under *the group model* will not fully adjust for the nonresponse bias, however, it should adjust for bias resulting from differential nonresponse across groups.

3.5 Analyzing the Bias due to Nonresponse in Joint Calibration Estimators

As in Särndal and Lundström (2005), the unconditional bias of the calibrated estimates can be evaluated jointly with respect to the sampling design $p(s)$ and the response distribution $q(r \mid s)$ as

$$Bias_{pq}\left(\hat{Y}_w\right) = E_p\left(E_q\left(\hat{Y}_w \mid s\right)\right) - Y = E_{pq}\left(\hat{Y}_w\right) - Y \tag{3.21}$$

Proposition 3.1

Under a univariate auxiliary variable $\mathbf{x}_k = 1$ for all $k \in U$, where the JCE can be written as in (3.12), and where response probabilities are $\Pr\left(k \in r_A \middle| s_A\right) = \phi_k^A$ and $\Pr\left(k \in r_B \middle| s_B\right) = \phi_k^B$ for samples A and B, respectively, the unconditional bias of JCE in (3.12), as derived in Appendix 3.1, can be approximately written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx N\left[\frac{\left(\sum_{U_A}\phi_k^A y_k + \sum_{U_B}\phi_k^B y_k\right)}{\left(\sum_{U_A}\phi_k^A + \sum_{U_B}\phi_k^B\right)} - \bar{Y}\right] \tag{3.22}$$

This means that bias is proportional to the difference between the response probabilities-weighted mean $\frac{\left(\sum_{U_A}\phi_k^A y_k + \sum_{U_B}\phi_k^B y_k\right)}{\left(\sum_{U_A}\phi_k^A + \sum_{U_B}\phi_k^B\right)}$ and the population mean $\bar{Y} = \sum_U y_k \big/ N$. Even if the response probabilities are constant throughout the population *U*, the bias in (3.22) does not vanish since

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx N\left[\frac{\left(\sum_{U_A} y_k + \sum_{U_B} y_k\right)}{N_A + N_B} - \bar{Y}\right] \tag{3.23}$$

The relative bias of JCE in (3.12), as derived in Appendix 3.1, can be written as

$$relbias\left(\hat{Y}_{JCE}\right) \approx \frac{Cov_{\phi_k^A;y_k} + Cov_{\phi_k^B;y_k}}{\bar{\phi}_U^{A,B}\,\bar{y}_U} \tag{3.24}$$

where $Cov_{\phi_k^A;y_k} = \frac{\sum_U \left(\phi_k^A - \overline{\phi}_U^A\right)\left(y_k - \overline{y}_U\right)}{N-1}$, $Cov_{\phi_k^B;y_k} = \frac{\sum_U \left(\phi_k^B - \overline{\phi}_U^B\right)\left(y_k - \overline{y}_U\right)}{N-1}$, $\overline{\phi}_U^A = \frac{\sum_U \phi_k^A}{N}$,

$\overline{\phi}_U^B = \frac{\sum_U \phi_k^B}{N}$ and $\overline{\phi}_U^{A,B} = \frac{\left(\sum_U \phi_k^A + \sum_U \phi_k^B\right)}{N}$. Even with constant response probabilities ϕ_k^A and ϕ_k^B throughout the two populations U_A and U_B, respectively, the covariance terms $Cov_{\phi_k^A;y_k}$ and $Cov_{\phi_k^B;y_k}$ do not equal zero.

Proposition 3.2

Under a continuous auxiliary variable $\mathbf{x}_k = x_k$ for all $k \in U$, where JCE can be written as in (3.15), the unconditional bias of JCE in (3.15), as derived in Appendix 3.2, can be approximately written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx X\left[\frac{\left(\sum_{U_A} \phi_k^A y_k + \sum_{U_B} \phi_k^B y_k\right)}{\left(\sum_{U_A} \phi_k^A x_k + \sum_{U_B} \phi_k^B x_k\right)} - \frac{Y}{X}\right] \tag{3.25}$$

This means that bias is proportional to the difference between the response probabilities-weighted mean $\frac{\left(\sum_{U_A} \phi_k^A y_k + \sum_{U_B} \phi_k^B y_k\right)}{\left(\sum_{U_A} \phi_k^A x_k + \sum_{U_B} \phi_k^B x_k\right)}$ and the population ratio $\frac{Y}{X}$. The relative bias of JCE in (3.12), as derived in Appendix 3.2, can be written as

$$relbias\left(\hat{Y}_{JCE}\right) \approx \frac{Cov_{\phi_k^A;y_k} + Cov_{\phi_k^B;y_k} + \overline{\phi}_U^{A,B}\overline{y}_U - \overline{\phi}_{xU}^{A,B}\overline{y}_U}{\overline{\phi}_{xU}^{A,B}\overline{y}_U} \tag{3.26}$$

where $\overline{\phi}_{xU}^{A,B} = \dfrac{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)}{\sum_U x_k}$.

Proposition 3.3

A general expression for the unconditional bias of JCE, as derived in Appendix 3.3, can be approximately written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = \sum_{U_{ab}} e_{k;\phi}^{A,B} - \sum_{U_A} e_{k;\phi}^{A,B}\left(1-\phi_k^A\right) - \sum_{U_B} e_{k;\phi}^{A,B}\left(1-\phi_k^B\right) \tag{3.27}$$

Where $e_{k;\phi}^{A,B} = \left(y_k - \mathbf{x}_k' B_{U;\phi}^{A,B}\right)$ and $\mathrm{B}_{U;\phi}^{A,B} = \left(\sum_{U_A} \phi_k^A \mathbf{x}_k y_k + \sum_{U_B} \phi_k^B \mathbf{x}_k y_k\right)\left(\sum_{U_A} \phi_k^A \mathbf{x}_k \mathbf{x}_k' + \sum_{U_B} \phi_k^B \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$.

Under full response when ϕ_k^A and ϕ_k^B are close to one,

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = \sum_{U_{ab}} e_k^{A,B} \tag{3.28}$$

This is consistent with bias under full response in proposition 2.1 in Chapter 2. Also, in the presence of the nonresponse, the bias of $\hat{Y}_{JCE}$ in (3.27) is independent of the sampling design used to draw s_A and s_B. However, the bias in (3.27) depends on the response distributions, $\Pr\left(k \in r_A | s_A\right)$ and $\Pr\left(k \in r_B | s_B\right)$, and their unknown response probabilities, ϕ_k^A and ϕ_k^B, for samples A and B, respectively. Proposition 3.3 emphasizes the need to identify powerful auxiliary variables that can predict study variable *y* and can identify response probabilities-homogeneous groups, in which ϕ_k^A and ϕ_k^B are homogeneous.

Corollary 3.3

Where a perfect linear relationship exists in the population between the study variable y_k and the auxiliary vector $\mathbf{x}_k$, as $y_k = \mathbf{x}_k' \beta$, for every $k \in U$, the bias of the JCE estimator, $\hat{Y}_{JCE}$, in (3.27) can be written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = \left(\sum_{U_{ab}} \mathbf{x}_k - \sum_{U_A} \mathbf{x}_k \left(1-\phi_k^A\right) - \sum_{U_B} \mathbf{x}_k \left(1-\phi_k^B\right)\right)\left(\mathrm{B}_U - \mathrm{B}_{U;\phi}^{A,B}\right) = 0 \qquad (3.29)$$

This is due the fact that when this perfect linear relationship between y_k and $\mathbf{x}_k$ exists, $\mathrm{B}_{U;\phi}^{A,B} = \mathrm{B}_U$. This shows that the bias of $\hat{Y}_{JCE}$ is a function of the difference between two regression vectors, $\mathrm{B}_{U;\phi}^{A,B}$ and B_U. This perfect relationship will not hold in practice. However, the bias in (3.27) will be reduced if the perfect linear relationship between y_k and $\mathbf{x}_k$ comes close to being attained. We should use auxiliary variables $\mathbf{x}_k$ such that the residuals $e_{k;\phi}^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_{U;\phi}^{A,B}\right)$ are small. This happens when $e_k = \left(y_k - \mathbf{x}_k' \mathrm{B}_U\right)$ are small. Using this set of auxiliary variables $\mathbf{x}_k$ guarantees reduced bias and variance of the JCE as a dual frame estimator in the presence of nonresponse.

3.6 Simulation study

Simulation studies were used to evaluate the performance of the JCE relative to the standard dual frame estimators under different nonresponse mechanisms. The simulated populations generated in Chapter 2 were used again in this chapter. The simulation factors are as the following

1. Sampling Designs

    a) Simple Sampling Design: simple random samples were selected from both frames.

    b) Complex Sampling Design: equally allocated stratified sample from frame A, and simple random sample from frame B.

2. Sample size: Equal allocation where $n_A = 1500$ and $n_B = 1500$.

3. Domain means: Large-differences domains' means where $\beta_a = 5$, $\beta_{ab} = 10$ and $\beta_b = 15$.

4. Correlation between y_{jk} and x_{jk}

    a) The population correlation coefficient is $\rho_{xy} = 0.40$.

    b) The population correlation coefficient is $\rho_{xy} = 0.60$.

    c) The population correlation coefficient is $\rho_{xy} = 0.80$.

5. Response mechanisms

    a) Simple Response Propensity Model (MCAR), where overall response rate is 30% (i.e., the response *R* has propensity $\text{pr}(R = 1) = .30$).

    b) Response Propensity by Auxiliary Variable *x* (MAR2), where

    $\text{pr}(R = 1 \mid x \le c) = .19$ and $\text{pr}(R = 1 \mid x > c) = .34$. *c* is the 1st quartile of *x*.

    c) Response Propensity by Frame (MAR3), where

$\text{pr}\left(R=1 \mid F=A\right)=.33$ and $\text{pr}\left(R=1 \mid F=B\right)=.25$.

d) Response Propensity by Frame and Auxiliary Variable x (MAR4), where

$\text{pr}\left(R=1 \mid F=A, x \le c\right)=.21$, $\text{pr}\left(R=1 \mid F=B, x \le c\right)=.16$,

$\text{pr}\left(R=1 \mid F=A, x > c\right)=.38$ and $\text{pr}\left(R=1 \mid F=B, x > c\right)=.28$.

e) Response Propensity by Design Domain (MAR5), where

$\text{pr}\left(R=1 \mid D=a\right)=.24$, $\text{pr}\left(R=1 \mid D=ab\left(A\right)\right)=.41$,

$\text{pr}\left(R=1 \mid D=b\right)=.14$ and $\text{pr}\left(R=1 \mid D=ab\left(B\right)\right)=.27$.

f) Response Propensity by Design Domain and Auxiliary Variable *x* (MAR6), where

$\text{pr}\left(R=1 \mid D=a, x \le c\right)=.14$, $\text{pr}\left(R=1 \mid D=ab\left(A\right), x \le c\right)=.29$,

$\text{pr}\left(R=1 \mid D=b, x \le c\right)=.09$, $\text{pr}\left(R=1 \mid D=ab\left(B\right), x \le c\right)=.19$,

$\text{pr}\left(R=1 \mid D=a, x > c\right)=.29$, $\text{pr}\left(R=1 \mid D=ab\left(A\right), x > c\right)=.44$,

$\text{pr}\left(R=1 \mid D=b, x > c\right)=.18$ and $\text{pr}\left(R=1 \mid D=ab\left(B\right), x > c\right)=.29$.

These sets of simulation factors combine to form 72 simulation studies, 36 simulation studies for each population model. One thousand replicates of initial samples of 3,000 cases were run for each study. To simulate a dual frame design, within each simulation replicate, two samples were drawn separately from both frames A and B. These samples were 'stacked' to form dual frame sample s. Conditional on the response mechanism, response sets r_A and r_B were realized using estimated response propensities applied within each frame sample.

3.7 Simulation Results

As in Chapter 2, only results for simple sampling design are discussed, since simulation results for complex sampling designs, in Appendix 3.4, show the same patterns of results. Generally, in the presence of nonresponse, biases in $\hat{Y}_{FWE}$ were present. Adding the calibration in the standard estimators, $\hat{Y}_{FWE}^{cal}$'s reduced the nonresponse bias of the estimator. Under the CLR model, in Table 3.1, the proposed $\hat{Y}_{JCE}$ estimator achieves relative biases comparable to the standard estimator calibrated versions, $\hat{Y}_{FWE}^{cal}$'s. Under the GLR model, in Figure 3.1 and Table 3.3, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ have higher relative biases than $\hat{Y}_{FWE.z}^{cal}$, $\hat{Y}_{FWE.x}^{cal}$ and $\hat{Y}_{FWE.xz}^{cal}$, respectively. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ results in reduced relative biases. Note that, for $\hat{Y}_{FWE}$, the relative biases in Table 3.1 and the RMSE's in Table 3.2 are almost the same. This implies that RMSE for $\hat{Y}_{FWE}$ is completely dominated by squared-bias. The same applies under the GLR model in Tables 3.3 and 3.4.

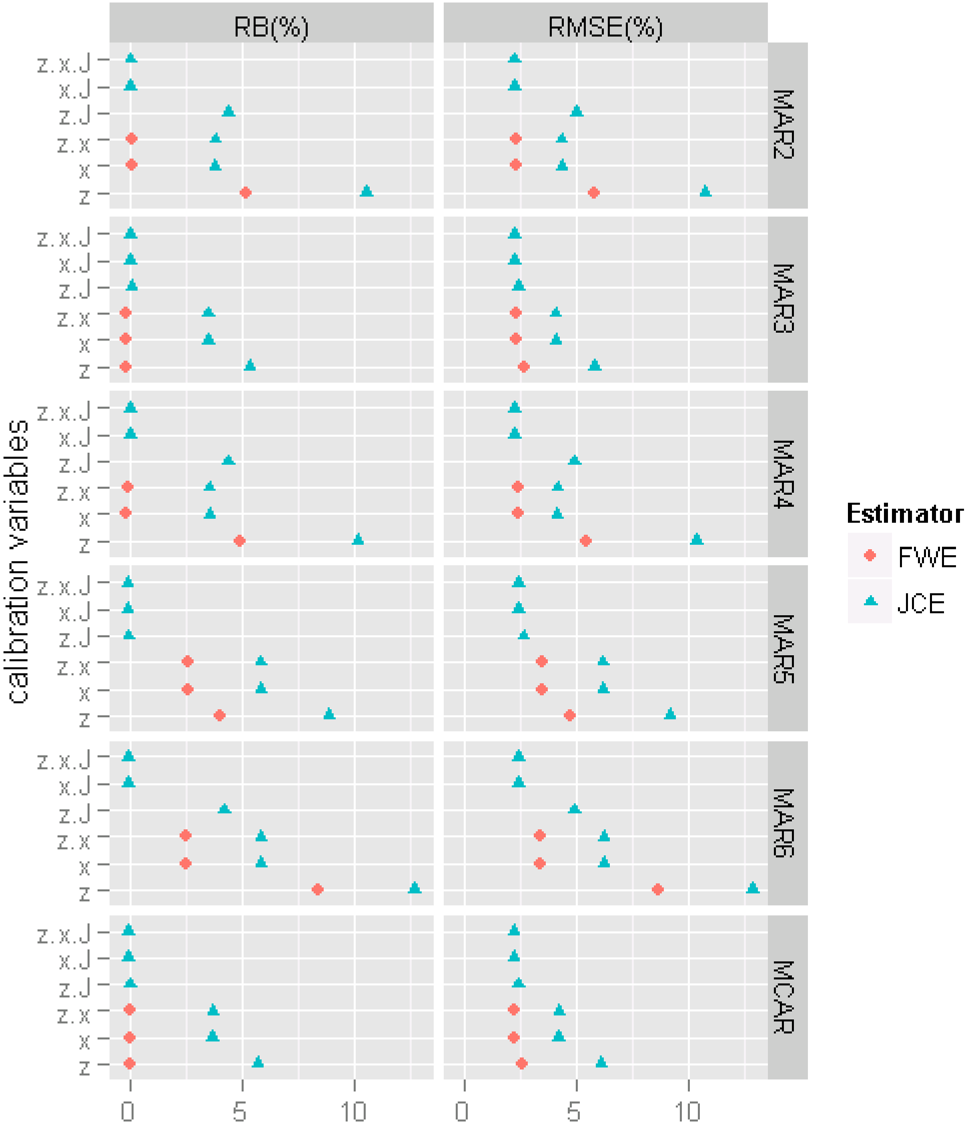
Under the CLR model in Table 3.1, adding the calibration by *z* in $\hat{Y}_{FWE.z}^{cal}$ resulted in lower relative nonresponse biases than the un-calibrated estimator $\hat{Y}_{FWE}$. Even lower relative nonresponse biases were achieved in the calibrated estimator $\hat{Y}_{FWE.x}^{cal}$ which uses *x* in the calibration step. Except for the *x*-dependent nonresponse mechanisms MAR2, MAR4 and MAR6, where the nonresponse mechanisms and the auxiliary variable *x* are dependent, calibrating with *z* in $\hat{Y}^{cal}_{FWE.z}$ was enough to adjust for the nonresponse errors. Also under the same *x*-dependent nonresponse mechanisms, the higher the correlation between *y* and *x*, the less successful is calibrating with *z,* in $\hat{Y}^{cal}_{FWE.z}$, in reducing the nonresponse bias. Calibrating with *x* in $\hat{Y}^{cal}_{FWE.x}$ was enough to adjust for the nonresponse errors under all nonresponse mechanisms and under all *y* and *x* correlation levels.

With regard to the JCE estimators in the presence of nonresponse errors, relative biases in $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.x}$ were comparable to relative biases in $\hat{Y}^{cal}_{FWE.z}$ and $\hat{Y}^{cal}_{FWE.x}$, respectively. Calibrating the 'stacked' samples directly by *x* in $\hat{Y}_{JCE.x}$ was enough to adjust for the nonresponse bias under all the nonresponse mechanisms and all *y* and *x* correlation levels. Except for the *x*-dependent nonresponse mechanisms, calibrating the 'stacked' samples directly by *z* in $\hat{Y}_{JCE.z}$ was enough to adjust for the nonresponse bias. Calibrating the 'stacked' samples directly by both *x* and *z* in $\hat{Y}_{JCE.xz}$ gave comparable results to $\hat{Y}_{JCE.x}$. Generally, the RMSE show similar patterns as the relative biases. However, RMSE's for $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.x}$ were slightly lower than RMSE's for $\hat{Y}^{cal}_{FWE.z}$ and $\hat{Y}^{cal}_{FWE.x}$, respectively.

Under the GLR model with 0.4 correlation level, in Figure 3.1, calibrating by *z* in $\hat{Y}^{cal}_{FWE.z}$ reduced the relative bias. However, biases for the case of *x*-dependent nonresponse mechanisms MAR2, MAR4 and MAR6, and *D*-dependent nonresponse mechanisms MAR5 and MAR6, remained. Calibrating by *x* in $\hat{Y}^{cal}_{FWE.x}$ reduced the relative bias for most of the nonresponse mechanisms. However, small relative biases for the *D*-dependent nonresponse mechanisms

MAR5 and MAR6 remained. The same applies under the other correlation levels, as in Table 3.3, however, the reduction in the relative biases is controlled by the correlation between *y* and *x*.

Figure 3.1: Simulation RB (%) and RMSE (%) for FWE and JCE estimators estimated from the GLR model population under simple sampling design and $\rho_{xy} = 0.40$



With regard to the proposed JCE estimators, calibrating the 'stacked' samples directly in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ reduced the error in comparison with the non-calibrated standard estimator $\hat{Y}_{FWE}$. However, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ have higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ reduced relative biases, and in contrast with $\hat{Y}^{cal}_{FWE.z}$, adding the strata totals to the calibration in $\hat{Y}_{JCE.zJ}$ resulted in reduced relative biases in one of the *D*-dependent nonresponse mechanisms, MAR5. In contrast with $\hat{Y}^{cal}_{FWE.x}$, adding the strata totals to the calibration in $\hat{Y}_{JCE.xJ}$ resulted in reduced relative biases under all the proposed nonresponse mechanisms. Again, under the GLR model, calibrating the 'stacked' samples directly by both *x* and *z* in $\hat{Y}_{JCE.xzJ}$ gave comparable results to $\hat{Y}_{JCE.xJ}$. The RMSE show same patterns as relative biases. However, RMSE's for $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ were slightly lower than RMSE's for $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively.

3.8 Discussion and Conclusion

In this chapter we addressed one of the desirable properties discussed in Chapter 1; the JCE was introduced as a dual frame estimator that is robust to the nonresponse errors conditional on using a powerful set of auxiliary variables. We investigated both the properties of the JCE as a dual frame estimator in the presence of nonresponse error and the implicit potentials of the JCE in attenuating the nonresponse bias under various nonresponse mechanisms. A general expression for the bias of the JCE estimator was derived. Expressions for Relative biases for different JCE estimators were derived, as well. The bias expression is composed of both nonresponse and dual frame estimation bias. These bias and relative bias expressions emphasize the need to identify powerful auxiliary variables that can predict study variable *y* and explain the nonresponse mechanism. Identifying these variables guarantees reduced bias and variance for the JCE as a dual frame estimator in the presence of nonresponse.

The performance of the JCE was explored empirically in the presence of nonresponse. When the auxiliary vector and the implicit calibration model more closely approximate the population model and the nonresponse mechanism, JCE yields almost unbiased dual frame estimates. This is consistent with Särndal and Lundström (2005) definition of powerful auxiliary vectors for reducing nonresponse, where the auxiliary vector should explain both the response propensity and the main study variables. The simulation results indicated that nonresponse can lead to biased dual frame estimates. Calibrating the FWE estimates may reduce the nonresponse bias. This reduction depends on using a set of strong auxiliary variables that explains the nonresponse mechanism. At the same time, the JCE results were comparable to the calibrated FWE estimates.

As derived theoretically, the correlation between the study variable and the response probabilities within each sample contribute to the increase of the estimates' relative biases. This is clear in the *x*-dependent nonresponse mechanisms where the relative biases were the highest among all the other mechanisms. In this case the correlation between the study variable and the response probabilities is due to the correlations between the auxiliary variable *x* and both the study variable and the response probabilities. Adding *x* to the calibration step either in FWE or in JCE is enough to adjust for the nonresponse bias under the *x*-dependent nonresponse mechanisms.

Finally, this chapter only addressed the nonresponse as one form of the non-sampling errors. More research is needed to explore the performance of the JCE in the presence of the other kinds of non-sampling errors. In the next chapter, we will address the effect of the measurement domain misclassification error on the JCE.

Table 3.1: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under simple sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
MCAR	$\rho_{xy} = 0.40$	-69.99	-0.01	-0.05	0.00	-0.02	0.00	-0.02
MAR2	$\rho_{xy} = 0.40$	-68.93	3.41	3.40	0.03	0.03	0.03	0.03
MAR3	$\rho_{xy} = 0.40$	-69.49	0.03	-0.01	0.03	0.01	0.03	0.02
MAR4	$\rho_{xy} = 0.40$	-68.44	3.62	3.63	0.02	0.04	0.02	0.04
MAR5	$\rho_{xy} = 0.40$	-71.96	0.00	-0.02	0.00	0.00	0.00	0.00
MAR6	$\rho_{xy} = 0.40$	-70.76	3.29	3.06	0.03	0.03	0.04	0.03
MCAR	$\rho_{xy} = 0.60$	-69.99	-0.03	-0.03	-0.01	0.00	-0.01	0.00
MAR2	$\rho_{xy} = 0.60$	-68.25	5.16	5.17	0.01	0.01	0.00	0.01
MAR3	$\rho_{xy} = 0.60$	-69.24	-0.02	-0.02	0.00	0.00	-0.01	0.00
MAR4	$\rho_{xy} = 0.60$	-67.65	5.17	5.14	0.02	0.02	0.02	0.02
MAR5	$\rho_{xy} = 0.60$	-71.93	-0.04	-0.04	-0.01	-0.01	-0.01	-0.01
MAR6	$\rho_{xy} = 0.60$	-70.05	4.68	4.30	-0.01	-0.02	-0.01	-0.02
MCAR	$\rho_{xy} = 0.80$	-69.98	0.00	-0.01	0.04	0.00	0.03	0.00
MAR2	$\rho_{xy} = 0.80$	-67.61	7.48	7.48	0.07	0.05	0.06	0.05
MAR3	$\rho_{xy} = 0.80$	-69.43	-0.03	-0.02	0.02	0.00	0.02	0.00
MAR4	$\rho_{xy} = 0.80$	-67.32	7.45	7.43	0.06	0.05	0.06	0.05
MAR5	$\rho_{xy} = 0.80$	-71.89	0.02	0.01	0.00	-0.02	0.00	-0.02
MAR6	$\rho_{xy} = 0.80$	-69.64	6.87	6.40	0.02	0.01	0.02	0.01

Table 3.2: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under simple sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
MCAR	$\rho_{xy} = 0.40$	69.99	2.03	1.91	1.87	1.75	1.88	1.76
MAR2	$\rho_{xy} = 0.40$	68.94	3.98	3.91	1.90	1.76	1.90	1.76
MAR3	$\rho_{xy} = 0.40$	69.50	2.05	1.94	1.88	1.75	1.88	1.76
MAR4	$\rho_{xy} = 0.40$	68.45	4.17	4.12	1.92	1.79	1.92	1.80
MAR5	$\rho_{xy} = 0.40$	71.97	2.10	1.96	1.93	1.80	1.93	1.80
MAR6	$\rho_{xy} = 0.40$	70.76	3.89	3.62	1.92	1.78	1.92	1.79
MCAR	$\rho_{xy} = 0.60$	70.00	2.11	1.93	1.71	1.56	1.72	1.56
MAR2	$\rho_{xy} = 0.60$	68.26	5.53	5.48	1.69	1.55	1.70	1.55
MAR3	$\rho_{xy} = 0.60$	69.25	2.07	1.90	1.69	1.55	1.70	1.55
MAR4	$\rho_{xy} = 0.60$	67.66	5.55	5.47	1.74	1.58	1.75	1.59
MAR5	$\rho_{xy} = 0.60$	71.94	2.11	1.94	1.74	1.60	1.75	1.60
MAR6	$\rho_{xy} = 0.60$	70.06	5.09	4.69	1.69	1.55	1.69	1.56
MCAR	$\rho_{xy} = 0.80$	69.99	2.21	2.04	1.33	1.24	1.33	1.24
MAR2	$\rho_{xy} = 0.80$	67.62	7.77	7.73	1.34	1.24	1.33	1.24
MAR3	$\rho_{xy} = 0.80$	69.44	2.22	2.05	1.33	1.24	1.32	1.24
MAR4	$\rho_{xy} = 0.80$	67.33	7.75	7.69	1.33	1.23	1.33	1.23
MAR5	$\rho_{xy} = 0.80$	71.89	2.18	2.04	1.34	1.25	1.34	1.25
MAR6	$\rho_{xy} = 0.80$	69.65	7.19	6.70	1.32	1.23	1.32	1.23

Table 3.3: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under simple sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
MCAR	$\rho_{xy} = 0.40$	-69.98	0.00	5.65	-0.07	3.65	-0.06	3.66	-0.03	-0.11	-0.10
MAR2	$\rho_{xy} = 0.40$	-68.47	5.18	10.52	0.03	3.76	0.03	3.77	4.36	-0.01	0.00
MAR3	$\rho_{xy} = 0.40$	-69.57	-0.26	5.30	-0.21	3.44	-0.20	3.44	0.01	-0.06	-0.06
MAR4	$\rho_{xy} = 0.40$	-68.38	4.86	10.12	-0.18	3.50	-0.17	3.51	4.33	-0.02	-0.02
MAR5	$\rho_{xy} = 0.40$	-70.77	4.00	8.86	2.57	5.82	2.58	5.82	-0.09	-0.16	-0.16
MAR6	$\rho_{xy} = 0.40$	-69.60	8.33	12.68	2.48	5.81	2.49	5.81	4.15	-0.10	-0.10
MCAR	$\rho_{xy} = 0.60$	-69.97	0.03	6.00	0.00	3.30	0.01	3.30	0.00	0.00	0.00
MAR2	$\rho_{xy} = 0.60$	-68.34	6.89	12.47	0.05	3.38	0.05	3.38	6.05	0.06	0.05
MAR3	$\rho_{xy} = 0.60$	-69.57	-0.27	5.63	-0.14	3.13	-0.14	3.13	0.05	0.06	0.06
MAR4	$\rho_{xy} = 0.60$	-67.77	6.82	12.31	-0.15	3.16	-0.15	3.16	6.28	0.07	0.07
MAR5	$\rho_{xy} = 0.60$	-70.94	4.30	9.44	2.39	5.29	2.40	5.28	-0.01	0.05	0.05
MAR6	$\rho_{xy} = 0.60$	-69.19	10.28	14.68	2.18	5.16	2.18	5.16	6.10	0.04	0.03
MCAR	$\rho_{xy} = 0.80$	-70.01	-0.10	5.95	-0.04	2.42	-0.04	2.42	-0.16	-0.07	-0.07
MAR2	$\rho_{xy} = 0.80$	-67.48	9.39	14.92	-0.07	2.42	-0.06	2.42	8.46	-0.10	-0.11
MAR3	$\rho_{xy} = 0.80$	-69.57	-0.46	5.52	-0.21	2.23	-0.21	2.23	-0.18	-0.09	-0.09
MAR4	$\rho_{xy} = 0.80$	-67.20	9.20	14.63	-0.21	2.27	-0.20	2.26	8.57	-0.09	-0.09
MAR5	$\rho_{xy} = 0.80$	-70.63	4.32	9.49	1.73	3.88	1.73	3.88	-0.18	-0.10	-0.10
MAR6	$\rho_{xy} = 0.80$	-68.61	12.49	16.72	1.56	3.79	1.57	3.78	8.11	-0.14	-0.14

Table 3.4: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under simple sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
MCAR	$\rho_{xy} = 0.40$	69.99	2.55	6.12	2.27	4.22	2.27	4.23	2.41	2.20	2.20
MAR2	$\rho_{xy} = 0.40$	68.48	5.77	10.77	2.35	4.35	2.35	4.35	4.98	2.25	2.25
MAR3	$\rho_{xy} = 0.40$	69.58	2.64	5.83	2.35	4.07	2.35	4.07	2.42	2.22	2.22
MAR4	$\rho_{xy} = 0.40$	68.39	5.49	10.39	2.39	4.15	2.39	4.16	4.94	2.26	2.26
MAR5	$\rho_{xy} = 0.40$	70.78	4.77	9.17	3.47	6.20	3.47	6.21	2.65	2.44	2.44
MAR6	$\rho_{xy} = 0.40$	69.61	8.71	12.90	3.43	6.22	3.44	6.22	4.91	2.43	2.43
MCAR	$\rho_{xy} = 0.60$	69.98	2.41	6.41	1.97	3.78	1.97	3.78	2.30	1.93	1.93
MAR2	$\rho_{xy} = 0.60$	68.35	7.30	12.67	1.96	3.85	1.96	3.85	6.47	1.90	1.90
MAR3	$\rho_{xy} = 0.60$	69.58	2.44	6.08	1.95	3.62	1.95	3.62	2.29	1.87	1.87
MAR4	$\rho_{xy} = 0.60$	67.78	7.23	12.51	1.98	3.66	1.98	3.66	6.69	1.91	1.91
MAR5	$\rho_{xy} = 0.60$	70.95	4.95	9.71	3.07	5.59	3.08	5.58	2.52	2.02	2.02
MAR6	$\rho_{xy} = 0.60$	69.20	10.56	14.86	2.97	5.50	2.97	5.49	6.60	2.09	2.09
MCAR	$\rho_{xy} = 0.80$	70.02	2.61	6.43	1.57	2.83	1.57	2.83	2.49	1.52	1.52
MAR2	$\rho_{xy} = 0.80$	67.49	9.71	15.09	1.58	2.84	1.58	2.84	8.79	1.53	1.53
MAR3	$\rho_{xy} = 0.80$	69.59	2.64	6.03	1.56	2.66	1.57	2.66	2.46	1.48	1.48
MAR4	$\rho_{xy} = 0.80$	67.21	9.54	14.81	1.62	2.72	1.62	2.72	8.89	1.54	1.54
MAR5	$\rho_{xy} = 0.80$	70.64	5.04	9.79	2.32	4.14	2.32	4.14	2.72	1.61	1.61
MAR6	$\rho_{xy} = 0.80$	68.62	12.74	16.89	2.24	4.07	2.24	4.07	8.53	1.67	1.67

Appendix 3.1: Proof of Proposition 3.1

$$\hat{Y}_{JCE} = N\left(\sum_{r_A} d_k y_k + \sum_{r_B} d_k y_k\right)\left(\sum_{r_A} d_k + \sum_{r_B} d_k\right)^{-1} \tag{3.30}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = E_{pq}\left(\hat{Y}_{JCE} - Y\right) \tag{3.31}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = N.E_{pq}\left(\frac{\left(\sum_{r_A} d_k y_k + \sum_{r_B} d_k y_k\right)}{\left(\sum_{r_A} d_k + \sum_{r_B} d_k\right)}\right) - Y \tag{3.32}$$

$$E_{pq}\left(\sum_{r_A} d_k y_k\right) = E_p\left(\sum_{s_A} \phi_k^A d_k y_k\right) = \sum_{U_A} \phi_k^A y_k \tag{3.33}$$

and

$$E_{pq}\left(\sum_{r_A} d_k\right) = E_p\left(\sum_{s_A} \phi_k^A d_k\right) = \sum_{U_A} \phi_k^A \tag{3.34}$$

Similarly, $E_{pq}\left(\sum_{r_B} d_k y_k\right) = \sum_{U_B} \phi_k^B y_k$ and $E_{pq}\left(\sum_{r_B} d_k\right) = \sum_{U_B} \phi_k^B$. So by replacing the numerator and the denominator in $\dfrac{\left(\sum_{r_A} d_k y_k + \sum_{r_B} d_k y_k\right)}{\left(\sum_{r_A} d_k + \sum_{r_B} d_k\right)}$ by their expected values, the unconditional bias can be approximately written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx N\left[\frac{\left(\sum_{U_A} \phi_k^A y_k + \sum_{U_B} \phi_k^B y_k\right)}{\left(\sum_{U_A} \phi_k^A + \sum_{U_B} \phi_k^B\right)} - \bar{Y}\right] \tag{3.35}$$

With regard to the relative bias, it can be written as

$$relbias\left(\hat{Y}_{JCE}\right) \approx \left[N \frac{\left(\sum_{U_A} \phi_k^A y_k + \sum_{U_B} \phi_k^B y_k\right)}{\left(\sum_{U_A} \phi_k^A + \sum_{U_B} \phi_k^B\right)\sum_U y_k} - 1 \right] \tag{3.36}$$

Since, $\phi_k^A = 0$ for all $k \in U_b$ and $\phi_k^B = 0$ for all $k \in U_a$, the relative bias can be written as

$$relbias\left(\hat{Y}_{JCE}\right) \approx \left[N \frac{\left(\sum_{U} \phi_k^A y_k + \sum_{U} \phi_k^B y_k\right)}{\left(\sum_{U} \phi_k^A + \sum_{U} \phi_k^B\right)\sum_U y_k} - 1 \right] \tag{3.37}$$

$$\approx \left[\frac{N\sum_U \phi_k^A y_k + N\sum_U \phi_k^B y_k + \sum_U \phi_k^A \sum_U y_k - \sum_U \phi_k^A \sum_U y_k + \sum_U \phi_k^B \sum_U y_k - \sum_U \phi_k^B \sum_U y_k}{\left(\sum_U \phi_k^A + \sum_U \phi_k^B\right)\sum_U y_k} - 1 \right] \tag{3.38}$$

$$\approx \left[\frac{N\left[\sum_U \left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right) + \sum_U \left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)\right] + \sum_U \phi_k^A \sum_U y_k + \sum_U \phi_k^B \sum_U y_k}{\left(\sum_U \phi_k^A + \sum_U \phi_k^B\right)\sum_U y_k} - 1 \right] \tag{3.39}$$

$$\approx \frac{N\left[\sum_U \left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right) + \sum_U \left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)\right]}{\left(\sum_U \phi_k^A + \sum_U \phi_k^B\right)\sum_U y_k} \tag{3.40}$$

Assume $\frac{N-1}{N} \approx 1$

$$relbias\left(\hat{Y}_{JCE}\right) \approx \frac{Cov_{\phi_k^A;y_k} + Cov_{\phi_k^B;y_k}}{\bar{\phi}_U^{A,B}\,\bar{y}_U} \tag{3.41}$$

where $Cov_{\phi_k^A;y_k} = \frac{\sum_U \left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right)}{N-1}$, $Cov_{\phi_k^B;y_k} = \frac{\sum_U \left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)}{N-1}$, $\bar{\phi}_U^A = \frac{\sum_U \phi_k^A}{N}$, $\bar{\phi}_U^B = \frac{\sum_U \phi_k^B}{N}$ and $\bar{\phi}_U^{A,B} = \frac{\left(\sum_U \phi_k^A + \sum_U \phi_k^B\right)}{N}$.

Appendix 3.2: Proof of Proposition 3.2

$$\hat{Y}_{JCE} = X\left(\sum\nolimits_{r_A} d_k y_k + \sum\nolimits_{r_B} d_k y_k\right)\left(\sum\nolimits_{r_A} d_k x_k + \sum\nolimits_{r_B} d_k x_k\right)^{-1} \tag{3.42}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) = X.E_{pq}\left(\frac{\sum\nolimits_{r_A} d_k y_k + \sum\nolimits_{r_B} d_k y_k}{\sum\nolimits_{r_A} d_k x_k + \sum\nolimits_{r_B} d_k x_k}\right) - Y \tag{3.43}$$

Where $E_{pq}\left(\sum\nolimits_{r_A} d_k x_k\right) = \sum\nolimits_{U_A} \phi_k^A x_k$ and $E_{pq}\left(\sum\nolimits_{r_B} d_k x_k\right) = \sum\nolimits_{U_B} \phi_k^B x_k$, the unconditional bias can be approximately written as

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx X\left[\frac{\left(\sum\nolimits_{U_A} \phi_k^A y_k + \sum\nolimits_{U_B} \phi_k^B y_k\right)}{\left(\sum\nolimits_{U_A} \phi_k^A x_k + \sum\nolimits_{U_B} \phi_k^B x_k\right)} - \frac{Y}{X}\right] \tag{3.44}$$

With regard to the relative bias, it can be written as

$$relbias\left(\hat{Y}_{JCE}\right) \approx \left[X\,\frac{\left(\sum\nolimits_{U_A} \phi_k^A y_k + \sum\nolimits_{U_B} \phi_k^B y_k\right)}{\left(\sum\nolimits_{U_A} \phi_k^A x_k + \sum\nolimits_{U_B} \phi_k^B x_k\right)\sum\nolimits_U y_k} - 1\right] \tag{3.45}$$

$$\approx \left[N\bar{X}\,\frac{\left(\sum\nolimits_U \phi_k^A y_k + \sum\nolimits_U \phi_k^B y_k\right)}{\left(\sum\nolimits_U \phi_k^A x_k + \sum\nolimits_U \phi_k^B x_k\right)\sum\nolimits_U y_k} - 1\right] \tag{3.46}$$

$$\approx \left[\frac{N\bar{X}\sum\nolimits_U \phi_k^A y_k + N\bar{X}\sum\nolimits_U \phi_k^B y_k + \bar{X}\sum\nolimits_U \phi_k^A \sum\nolimits_U y_k - \bar{X}\sum\nolimits_U \phi_k^A \sum\nolimits_U y_k + \bar{X}\sum\nolimits_U \phi_k^B \sum\nolimits_U y_k - \bar{X}\sum\nolimits_U \phi_k^B \sum\nolimits_U y_k}{\left(\sum\nolimits_U \phi_k^A x_k + \sum\nolimits_U \phi_k^B x_k\right)\sum\nolimits_U y_k} - 1\right] \tag{3.47}$$

$$\approx \left[\frac{N\bar{X}\left[\sum\nolimits_U\left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right) + \sum\nolimits_U\left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)\right] + \left[\sum\nolimits_U \phi_k^A + \sum\nolimits_U \phi_k^B\right]\bar{X}\sum\nolimits_U y_k}{\left(\sum\nolimits_U \phi_k^A x_k + \sum\nolimits_U \phi_k^B x_k\right)\sum\nolimits_U y_k} - 1\right] \tag{3.48}$$

$$\approx \left[\frac{N\bar{X}\left[\sum_U \left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right) + \sum_U \left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)\right]}{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)\sum_U y_k} + \frac{\left[\sum_U \phi_k^A + \sum_U \phi_k^B\right]\bar{X}}{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)} - 1 \right] \tag{3.49}$$

$$\approx \left[\frac{\left[\sum_U \left(\phi_k^A - \bar{\phi}_U^A\right)\left(y_k - \bar{y}_U\right) + \sum_U \left(\phi_k^B - \bar{\phi}_U^B\right)\left(y_k - \bar{y}_U\right)\right]}{\frac{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)}{\sum_U x_k}\sum_U y_k} + \frac{\left[\sum_U \phi_k^A + \sum_U \phi_k^B\right]\Big/ N}{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)\Big/\sum_U x_k} - 1 \right] \tag{3.50}$$

Assume $\frac{N-1}{N} \approx 1$

$$relbias\left(\hat{Y}_{JCE}\right) \approx \frac{Cov_{\phi_k^A;y_k} + Cov_{\phi_k^B;y_k}}{\bar{\phi}_{xU}^{A,B}\,\bar{y}_U} + \frac{\bar{\phi}_U^{A,B}}{\bar{\phi}_{xU}^{A,B}} - 1 \tag{3.51}$$

$$\approx \frac{Cov_{\phi_k^A;y_k} + Cov_{\phi_k^B;y_k} + \bar{\phi}_U^{A,B}\,\bar{y}_U - \bar{\phi}_{xU}^{A,B}\,\bar{y}_U}{\bar{\phi}_{xU}^{A,B}\,\bar{y}_U} \tag{3.52}$$

where

$$\bar{\phi}_{xU}^{A,B} = \frac{\left(\sum_U \phi_k^A x_k + \sum_U \phi_k^B x_k\right)}{\sum_U x_k} \tag{3.53}$$

Appendix 3.3: Proof of Proposition 3.3

$$\begin{aligned}\hat{Y}_{JCE} &= \sum\nolimits_{r_A} d_k y_k + \sum\nolimits_{r_B} d_k y_k \\ &+ \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{r_A} d_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}'_k\right)\right)\left(\sum\nolimits_{r_A} d_k \mathbf{x}_k y_k + \sum\nolimits_{r_B} d_k \mathbf{x}_k y_k\right)\left(\sum\nolimits_{r_A} d_k \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\end{aligned} \tag{3.54}$$

$$\begin{aligned}\hat{Y}_{JCE} &= \sum\nolimits_{r_A} d_k y_k + \sum\nolimits_{r_B} d_k y_k + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{r_A} d_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}'_k\right)\right)\mathrm{B}_{U;\phi}^{A,B} \\ &+ \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{r_A} d_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}'_k\right)\right)\hat{B}_r^{A,B} - \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{r_A} d_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}'_k\right)\right)\mathrm{B}_{U;\phi}^{A,B}\end{aligned} \tag{3.55}$$

where

$$\hat{B}_r^{A,B} = \left(\sum\nolimits_{r_A} d_k \mathbf{x}_k y_k + \sum\nolimits_{r_B} d_k \mathbf{x}_k y_k\right)\left(\sum\nolimits_{r_A} d_k \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \tag{3.56}$$

and

$$\mathrm{B}_{U;\phi}^{A,B} = \left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}_k y_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}_k y_k\right)\left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \tag{3.57}$$

$$\begin{aligned}E_{pq}\left(\hat{Y}_{JCE}\right) &= \sum\nolimits_{U_A} \phi_k^A y_k + \sum\nolimits_{U_B} \phi_k^B y_k + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}'_k\right)\right)\mathrm{B}_{U;\phi}^{A,B} \\ &+ \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}'_k\right)\right)E_{pq}\left(\hat{B}_r^{A,B} - \mathrm{B}_{U;\phi}^{A,B}\right)\end{aligned} \tag{3.58}$$

$$E_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_A} \phi_k^A y_k + \sum\nolimits_{U_B} \phi_k^B y_k + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}'_k\right)\right)\mathrm{B}_{U;\phi}^{A,B} \tag{3.59}$$

Since

$$\begin{aligned}&\text{plim } \hat{B}_r^{A,B} = \text{plim } \left\{\left(\sum\nolimits_{r_A} d_k^A \mathbf{x}_k y_k + \sum\nolimits_{r_B} d_k^B \mathbf{x}_k y_k\right)\left(\sum\nolimits_{r_A} d_k^A \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{r_B} d_k^B \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\right\} = \\ &\left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}_k y_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}_k y_k\right)\left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}_k \mathbf{x}'_k\right)^{-1} = \mathrm{B}_{U;\phi}^{A,B}\end{aligned} \tag{3.60}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_A} \phi_k^A y_k + \sum\nolimits_{U_B} \phi_k^B y_k + \left(\sum\nolimits_{U} x_k - \left(\sum\nolimits_{U_A} \phi_k^A \mathbf{x}'_k + \sum\nolimits_{U_B} \phi_k^B \mathbf{x}'_k\right)\right)\mathrm{B}_{U;\phi}^{A,B} - \sum\nolimits_{U} y_k \tag{3.61}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_A} \phi_k^A \left(y_k - \mathbf{x}_k' \mathrm{B}_{U;\phi}^{A,B}\right) + \sum\nolimits_{U_B} \phi_k^B \left(y_k - \mathbf{x}_k' \mathrm{B}_{U;\phi}^{A,B}\right) - \sum\nolimits_{U} \left(y_k - \mathbf{x}_k' \mathrm{B}_{U;\phi}^{A,B}\right) \tag{3.62}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_A} \phi_k^A e_{k;\phi}^{A,B} + \sum\nolimits_{U_B} \phi_k^B e_{k;\phi}^{A,B} - \sum\nolimits_{U} e_{k;\phi}^{A,B} \tag{3.63}$$

where

$$e_{k;\phi}^{A,B} = \left(y_k - \mathbf{x}_k' \mathrm{B}_{U;\phi}^{A,B}\right) \tag{3.64}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_A} \phi_k^A e_{k;\phi}^{A,B} + \sum\nolimits_{U_B} \phi_k^B e_{k;\phi}^{A,B} - \sum\nolimits_{U_A} e_{k;\phi}^{A,B} - \sum\nolimits_{U_B} e_{k;\phi}^{A,B} + \sum\nolimits_{U_{ab}} e_{k;\phi}^{A,B} \tag{3.65}$$

$$Bias_{pq}\left(\hat{Y}_{JCE}\right) \approx \sum\nolimits_{U_{ab}} e_{k;\phi}^{A,B} - \sum\nolimits_{U_A} e_{k;\phi}^{A,B} \left(1 - \phi_k^A\right) - \sum\nolimits_{U_B} e_{k;\phi}^{A,B} \left(1 - \phi_k^B\right) \tag{3.66}$$

Appendix 3.4: Results for Complex Sampling Designs

Table 3.5: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under complex sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
MCAR	$\rho_{xy} = 0.40$	-69.98	0.04	0.00	0.05	0.03	0.05	0.03
MAR2	$\rho_{xy} = 0.40$	-68.98	3.67	3.65	0.06	0.06	0.05	0.05
MAR3	$\rho_{xy} = 0.40$	-69.54	0.05	0.01	0.04	0.03	0.04	0.03
MAR4	$\rho_{xy} = 0.40$	-68.32	3.58	3.55	0.09	0.09	0.08	0.09
MAR5	$\rho_{xy} = 0.40$	-71.66	-0.01	-0.04	0.01	0.01	0.00	0.00
MAR6	$\rho_{xy} = 0.40$	-70.80	3.35	3.10	0.02	0.03	0.02	0.02
MCAR	$\rho_{xy} = 0.60$	-70.02	-0.10	-0.09	-0.08	-0.07	-0.09	-0.07
MAR2	$\rho_{xy} = 0.60$	-68.53	4.93	4.95	-0.09	-0.07	-0.09	-0.07
MAR3	$\rho_{xy} = 0.60$	-69.56	-0.05	-0.05	-0.06	-0.05	-0.06	-0.06
MAR4	$\rho_{xy} = 0.60$	-67.72	4.99	4.98	-0.07	-0.06	-0.08	-0.06
MAR5	$\rho_{xy} = 0.60$	-71.93	-0.02	-0.01	-0.03	-0.02	-0.03	-0.02
MAR6	$\rho_{xy} = 0.60$	-70.19	4.85	4.51	-0.06	-0.04	-0.06	-0.04
MCAR	$\rho_{xy} = 0.80$	-69.98	0.05	0.07	0.02	-0.01	0.02	-0.01
MAR2	$\rho_{xy} = 0.80$	-67.50	7.27	7.32	0.02	0.01	0.02	0.01
MAR3	$\rho_{xy} = 0.80$	-69.53	0.01	0.04	-0.02	-0.03	-0.02	-0.03
MAR4	$\rho_{xy} = 0.80$	-67.22	7.31	7.32	0.02	0.02	0.02	0.02
MAR5	$\rho_{xy} = 0.80$	-71.69	0.04	0.08	0.00	-0.01	0.00	-0.01
MAR6	$\rho_{xy} = 0.80$	-69.59	6.95	6.57	0.03	0.03	0.03	0.03

Table 3.6: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under complex sampling design and 30 % response rate.

Non-response	ρxy	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
MCAR	$\rho_{xy} = 0.40$	69.99	2.34	2.18	2.12	1.98	2.13	1.98
MAR2	$\rho_{xy} = 0.40$	68.99	4.30	4.21	2.14	1.99	2.14	1.99
MAR3	$\rho_{xy} = 0.40$	69.55	2.33	2.20	2.15	2.01	2.15	2.02
MAR4	$\rho_{xy} = 0.40$	68.33	4.26	4.15	2.17	2.02	2.17	2.02
MAR5	$\rho_{xy} = 0.40$	71.67	2.36	2.23	2.14	2.02	2.15	2.02
MAR6	$\rho_{xy} = 0.40$	70.81	4.10	3.83	2.21	2.08	2.21	2.08
MCAR	$\rho_{xy} = 0.60$	70.03	2.21	2.08	1.72	1.64	1.72	1.64
MAR2	$\rho_{xy} = 0.60$	68.54	5.37	5.34	1.74	1.64	1.74	1.65
MAR3	$\rho_{xy} = 0.60$	69.57	2.23	2.12	1.75	1.66	1.75	1.66
MAR4	$\rho_{xy} = 0.60$	67.73	5.43	5.38	1.78	1.68	1.78	1.69
MAR5	$\rho_{xy} = 0.60$	71.94	2.21	2.11	1.76	1.68	1.76	1.69
MAR6	$\rho_{xy} = 0.60$	70.20	5.31	4.96	1.77	1.71	1.78	1.71
MCAR	$\rho_{xy} = 0.80$	69.99	2.30	2.14	1.41	1.32	1.41	1.32
MAR2	$\rho_{xy} = 0.80$	67.51	7.59	7.59	1.40	1.31	1.39	1.31
MAR3	$\rho_{xy} = 0.80$	69.54	2.32	2.17	1.42	1.34	1.42	1.34
MAR4	$\rho_{xy} = 0.80$	67.23	7.63	7.60	1.42	1.35	1.43	1.35
MAR5	$\rho_{xy} = 0.80$	71.70	2.35	2.20	1.42	1.35	1.42	1.35
MAR6	$\rho_{xy} = 0.80$	69.60	7.31	6.90	1.42	1.35	1.42	1.34

Table 3.7: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under complex sampling design and 30 % response rate.

Non-response	ρ_{xy}	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
MCAR	$\rho_{xy} = 0.40$	-70.00	-0.02	5.70	-0.02	3.76	-0.02	3.77	0.01	-0.02	-0.01
MAR2	$\rho_{xy} = 0.40$	-68.87	5.09	10.52	-0.02	3.80	-0.01	3.80	4.31	-0.01	-0.01
MAR3	$\rho_{xy} = 0.40$	-69.73	-0.34	5.31	-0.23	3.50	-0.23	3.51	-0.02	-0.04	-0.04
MAR4	$\rho_{xy} = 0.40$	-68.52	4.78	10.12	-0.22	3.55	-0.21	3.56	4.30	0.00	0.00
MAR5	$\rho_{xy} = 0.40$	-70.59	4.11	9.02	2.71	6.00	2.72	6.00	0.02	-0.03	-0.02
MAR6	$\rho_{xy} = 0.40$	-69.56	8.29	12.74	2.49	5.90	2.50	5.91	4.19	-0.02	-0.01
MCAR	$\rho_{xy} = 0.60$	-70.02	-0.08	5.87	-0.10	3.21	-0.09	3.21	-0.12	-0.10	-0.10
MAR2	$\rho_{xy} = 0.60$	-68.37	7.17	12.71	-0.08	3.26	-0.08	3.25	6.28	-0.06	-0.07
MAR3	$\rho_{xy} = 0.60$	-69.68	-0.42	5.46	-0.28	2.99	-0.28	2.99	-0.09	-0.06	-0.06
MAR4	$\rho_{xy} = 0.60$	-67.77	6.72	12.20	-0.23	3.08	-0.23	3.07	6.20	-0.01	-0.01
MAR5	$\rho_{xy} = 0.60$	-70.61	4.11	9.25	2.24	5.15	2.24	5.15	-0.15	-0.07	-0.08
MAR6	$\rho_{xy} = 0.60$	-69.24	10.36	14.71	2.14	5.13	2.14	5.13	6.13	-0.03	-0.03
MCAR	$\rho_{xy} = 0.80$	-70.00	-0.02	6.04	0.05	2.52	0.05	2.52	-0.10	0.02	0.02
MAR2	$\rho_{xy} = 0.80$	-67.45	9.19	14.74	0.00	2.49	0.00	2.49	8.25	-0.04	-0.04
MAR3	$\rho_{xy} = 0.80$	-69.56	-0.36	5.63	-0.13	2.33	-0.13	2.33	-0.12	-0.02	-0.02
MAR4	$\rho_{xy} = 0.80$	-66.85	9.05	14.50	-0.16	2.32	-0.16	2.32	8.43	-0.04	-0.04
MAR5	$\rho_{xy} = 0.80$	-70.60	4.21	9.46	1.77	3.94	1.77	3.94	-0.20	-0.03	-0.03
MAR6	$\rho_{xy} = 0.80$	-68.48	12.47	16.76	1.70	3.92	1.70	3.92	8.17	-0.02	-0.02

Table 3.8: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under complex sampling design and 30 % response rate.

Non-response	ρxy	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$
MCAR	$\rho_{xy} = 0.40$	70.01	2.63	6.23	2.40	4.39	2.39	4.39	2.49	2.29	2.28
MAR2	$\rho_{xy} = 0.40$	68.88	5.72	10.81	2.45	4.46	2.44	4.46	4.97	2.32	2.32
MAR3	$\rho_{xy} = 0.40$	69.74	2.73	5.92	2.46	4.22	2.45	4.22	2.56	2.33	2.32
MAR4	$\rho_{xy} = 0.40$	68.53	5.48	10.44	2.48	4.28	2.47	4.29	4.99	2.36	2.35
MAR5	$\rho_{xy} = 0.40$	70.59	4.89	9.38	3.63	6.44	3.63	6.44	2.69	2.46	2.45
MAR6	$\rho_{xy} = 0.40$	69.57	8.73	13.01	3.55	6.39	3.55	6.39	5.00	2.54	2.53
MCAR	$\rho_{xy} = 0.60$	70.03	2.56	6.34	2.06	3.76	2.06	3.75	2.36	1.93	1.93
MAR2	$\rho_{xy} = 0.60$	68.38	7.59	12.92	2.09	3.81	2.09	3.81	6.70	1.96	1.96
MAR3	$\rho_{xy} = 0.60$	69.70	2.61	5.99	2.08	3.58	2.08	3.57	2.37	1.92	1.92
MAR4	$\rho_{xy} = 0.60$	67.78	7.17	12.43	2.11	3.67	2.11	3.67	6.62	1.95	1.95
MAR5	$\rho_{xy} = 0.60$	70.62	4.86	9.58	3.04	5.50	3.04	5.50	2.53	2.03	2.03
MAR6	$\rho_{xy} = 0.60$	69.25	10.67	14.92	3.02	5.52	3.02	5.51	6.64	2.08	2.08
MCAR	$\rho_{xy} = 0.80$	70.01	2.71	6.57	1.63	2.96	1.63	2.96	2.53	1.55	1.55
MAR2	$\rho_{xy} = 0.80$	67.47	9.55	14.94	1.67	2.96	1.67	2.96	8.61	1.58	1.58
MAR3	$\rho_{xy} = 0.80$	69.57	2.76	6.21	1.65	2.80	1.65	2.80	2.55	1.55	1.55
MAR4	$\rho_{xy} = 0.80$	66.86	9.42	14.71	1.69	2.82	1.69	2.82	8.77	1.58	1.58
MAR5	$\rho_{xy} = 0.80$	70.61	5.03	9.81	2.42	4.24	2.42	4.24	2.69	1.65	1.65
MAR6	$\rho_{xy} = 0.80$	68.49	12.76	16.96	2.42	4.25	2.42	4.25	8.59	1.70	1.70

Chapter 4

Joint Calibration Estimator in the Presence of Domain Misclassification

4.1 Introduction

After exploring the performance of JCE in the presence of nonresponse in Chapter 3, in this chapter, we are interested in exploring the performance of JCE in comparison with the standard dual frame estimators in the presence of domain misclassification for sample units. In this chapter, the domain misclassification problem is discussed in Section 4.2. The misclassification bias for the standard dual frame estimators is derived in Section 4.3. The performance of JCE in comparison with the standard FWE dual frame estimator in the presence of the misclassification and nonresponse errors is explored by a simulation study described in Section 4.4. The simulation results and findings are presented and discussed in Sections 4.5 and 4.6.

4.2 Domain Misclassification in Dual Frame Design

In Chapters 2 and 3, the Joint Calibration Estimator (JCE) was introduced as a new dual frame estimator that requires a simpler set of requirements than standard dual frame estimators while achieving comparable efficiency. Standard dual frame estimators depend on identifying the design domains during the data collection. Consequently, the performance of these estimators is sensitive to the errors in measuring the domain membership, or the simple multiplicity information (Mecatti, 2007). The multiplicity information problems include (i) Domain misclassification; this problem will be discussed in detail in this chapter, (ii) Item missing nonresponse in the multiplicity information, where the standard imputation can be used to impute the missing multiplicity information (Rubin, 1987) and (iii) Unknown multiplicity information, where for some groups the multiplicity information cannot be identified. For example, Zero-banks are 100-series banks with no listed residential numbers (Casady & Lepkowski, 1993). The list-assisted RDD designs do not implicitly cover phone numbers in the Zero-banks, but the ABS frame does cover this domain. Where many dual frame designs can be generated from combining RDD landline telephone sample and an ABS sample, in some designs, the Zero-banks domain should be explicitly identified in the ABS sample for sake of the unbiased dual frame estimation. However, practically speaking, identifying whether the sample case belongs to a zero-bank or not is not an easy task. Thus zero-banks cases will be forcibly embedded within the landline households.

Back to the domain misclassification problem, it is uncommon to have access to the domain membership information before conducting the survey (e.g., from properties or actual matching of frames). Therefore, this information should be obtained during the data collection. For example information about landline telephone service should be obtained in the area-landline dual frame surveys (Lepkowski & Groves, 1986); even more detailed information about the landline and cell phone services should be obtained in the landline-cell dual frame telephone surveys (Brick et al., 2006; Kennedy, 2007). Collecting this information could be burdensome for some respondents and could lead to more unit non-response. It is even worse when dealing with rare populations such as persons with a rare disease or for elusive or hidden populations such as the homeless, illegal immigrants or drug consumers (Lepkowski, 1991; Mecatti, 2007; Sudman & Kalton, 1986). For example, in The National Incidence Study of Child Abuse and Neglect, a dual frame design combines a list frame of all maltreated children investigated by Child Protective Services agencies and another sample frame compiled from reports of maltreated children provided by sources such as the police and school staff. Due to the lack of a list frame for the second frame, identifying the domain membership or the multiplicity information was problematic (Clark, Winglee, & Liu, 2007).

Beside the knowledge of the domain membership for every sampled unit, ideally, such information should be free from reporting or measurement errors, but this is not typically the case (Lohr & Rao, 2006). The correct classification of the sampled units into the domains in each frame is required to apply either the optimal or practical dual frame estimators as discussed in Chapter 1. The performance of the optimal dual frame estimator is dependent on the correct classification assumption. In practice, achieving the correct classification for all cases is almost impossible because, as any other study variable, the domain membership variable could be affected by the measurement or the reporting error. Therefore, the sampled units could be misclassified into the wrong domain. Misclassification happens when a sample unit is classified into the wrong design domain, such as when in RDD-cell phone dual frame surveys, households owning both landline and cell phone are misclassified as landline only households. In agriculture dual frame surveys, domain misclassification occurs if a farm sampled in the area frame is incorrectly classified with respect to its list frame membership (Lesser & Kalsbeek, 1999). It is even more challenging in longitudinal dual frame surveys (Lu & Lohr, 2010). Generally, it is difficult to identify misclassified units, and to estimate the misclassification rate. This means that the optimal dual frame estimators could have less than optimal performance (Lohr, 2011; Lohr & Rao, 2006).

As discussed in Chapter 1, Meccati (2007) introduced the Multiplicity Estimator (ME) (1.17), which depends on partial multiplicity information M_k , the number of the frames that case *k* belongs to, in order to adjust for the multiplicity and combine the different samples. Although the ME estimator somewhat relaxes the burden of collecting full multiplicity information, it is sensitive to the domain membership misclassification. Unlike the standard dual frame estimators, identifying the design domains is not necessarily required for the JCE. Thus, the JCE should be robust to multiplicity information problems such as missing multiplicity information or misclassification. In the presence of the multiplicity problems, the joint calibration approach tends to have higher efficiency than the standard dual frame estimators. In the next section, a misclassification bias expression is derived to help identify misclassification bias components. This formulation will enhance our understanding of the effect of the misclassification error on the standard dual frame estimates.

4.3 Analyzing the Bias due to Domain Misclassification in the Standard Dual Frame Estimators

In this section, the analytic bias due to domain misclassification is derived. This bias affects the standard dual frame estimators. However, it does not affect the JCE, which does not necessarily require any domain membership information for its general application. In the presence of domain misclassification and where s_{mis} is the domain-misclassified sample *s*, the unconditional bias of the standard dual frame estimators in (1.4), $\hat{Y}_{mis}$, can be evaluated jointly with respect to the sampling design $p(s)$ and the conditional misclassification distribution $q(s_{mis} \mid s)$ as

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E_p\left(E_q\left(\hat{Y}_{mis} \mid s\right)\right) - Y = E_{pq}\left(\hat{Y}_{mis}\right) - Y \tag{4.1}$$

Since the domain misclassification is more likely to occur in certain directions (Lohr, 2011), the domain misclassification can be classified as a one-way or two-way misclassification. Under the one-way misclassification, the misclassification problem can occur only from the overlapping to the non-overlapping domains, One-Way Overlapping Misclassification (OWOM), where the sample cases in the overlapping domains, s_{ab}^{A} and s_{ab}^{B}, could be misclassified into the non-overlapping domains, s_a and s_b. Another one-way misclassification mechanism happens when the misclassification problem occurs only from the non-overlapping to the overlapping domains, One-Way Non-overlapping Misclassification (OWNM), where the sample cases in the non-overlapping domains, s_a and s_b, could be misclassified into the overlapping domains, s_{ab}^{A} and s_{ab}^{B}. In the two-way misclassification (TWM), the sample cases either in the overlapping domain or in the non-overlapping domains could be misclassified into the wrong domains.

In the following propositions, the domain misclassification bias in the standard dual frame estimators is derived. In propositions 4.1 and 4.2, the analytic bias for the one-way misclassifications, OWOM and OWNM, are derived. In proposition 4.3, the analytic bias for the two-way misclassification, TWM, is derived.

Proposition 4.1

Under the one-way OWOM misclassification, where the sample cases in the overlapping domains, s_{ab}^{A} and s_{ab}^{B}, could be misclassified into the non-overlapping domains, s_a and s_b, the estimated total *Y* in (1.4) can be written as

$$\hat{Y}_{mis} = \sum\nolimits_{s_c} d_k y_k + \sum\nolimits_{s_{ab}} \left(1 - I_k^{ab,c}\right) m_k d_k y_k + \sum\nolimits_{s_{ab}} I_k^{ab,c} d_k y_k \tag{4.2}$$

where $d_k = \left(d_k^A, d_k^B\right)$, $m_k = \left(m_k^A, m_k^B\right)$, $s_c = \left(s_a, s_b\right)$, $s_{ab} = \left(s_{ab}^A, s_{ab}^B\right)$ and $I_k^{ab,c}$ is a misclassification indicator for observation *k* from the overlapping domains s_{ab}^{A} and s_{ab}^{B} misclassified into non-overlapping domains s_a and s_b, respectively.

Under this misclassification pattern, the unconditional bias resulting from domain misclassification, as derived in Appendix 4.1, can be written as

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = \gamma^{ab,c} Y_{ab} \tag{4.3}$$

where $\gamma^{ab,c}$ is the probability of misclassification from the overlapping to the non-overlapping domains. This means that the value of the misclassification probability $\gamma^{ab,c}$ and the population total *Y* for overlapping domains Y_{ab} determine the magnitude of the bias resulting from misclassification.

Proposition 4.2

Under the one-way OWNM misclassification, where the sample cases in the non-overlapping domains, s_a and s_b, could be misclassified into the overlapping domains, s_{ab}^A and s_{ab}^B, the estimated total *Y* in (1.4) can be written as

$$\hat{Y}_{mis} = \sum\nolimits_{s_{ab}} m_k d_k y_k + \sum\nolimits_{s_c} \left(1 - I_k^{c,ab}\right) d_k y_k + \sum\nolimits_{s_c} I_k^{c,ab} m_k d_k y_k \tag{4.4}$$

where $I_k^{c,ab}$ is a misclassification indicator for observation *k* from the non-overlapping domains s_a and s_b misclassified into overlapping domains s_{ab}^A and s_{ab}^B, respectively.

Under this misclassification pattern, the unconditional bias resulting from domain misclassification, as derived in Appendix 4.2, can be written as

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = \gamma^{c,ab}\left(\left(\theta - 1\right)Y_a - \theta Y_b\right) \tag{4.5}$$

Where $\theta \in \left[0,1\right]$ is the dual frame estimation composite factor and $\gamma^{c,ab}$ is the probability of misclassification from the non-overlapping to the overlapping domains. This means that the misclassification probability $\gamma^{c,ab}$ and a weighted average of population totals for non-overlapping domains $\left(\left(\theta - 1\right)Y_a - \theta Y_b\right)$ determine the magnitude of bias resulting from domain misclassification.

Proposition 4.3

Under the two-way TWM misclassification, where the sample cases either in the overlapping domains or in the non-overlapping domains could be misclassified into the wrong domains, the net unconditional bias resulting from misclassification, can be written as

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=\gamma^{ab,c}Y_{ab}-\gamma^{c,ab}\left(\left(1-\theta\right)Y_a+\theta Y_b\right) \tag{4.6}$$

These misclassification biases in (4.3), (4.5) and (4.6) assume that the misclassification indicators are identically distributed and that $I_k^{c,ab}$ and $I_k^{ab,c}$ are Bernoulli random variables with parameters $\gamma^{c,ab}$ and $\gamma^{ab,c}$, respectively. In the following proposition, a general expression for the misclassification bias is derived.

Proposition 4.4

A general expression for the unconditional bias resulting from the two-way TWM misclassification that assumes that each element *k* in the overlapping domain has a misclassification probability, $E\left(I_k^{ab,c}\right)=\gamma_k^{ab,c}$ and each element *k* in the non-overlapping domains has a misclassification probability $E\left(I_k^{c,ab}\right)=\gamma_k^{c,ab}$, as derived in Appendix 4.3, can be written as

$$\begin{aligned}Bias_{pq}\left(\hat{Y}_{mis}\right)&=N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)+\overline{\gamma}^{ab,c}\overline{Y}_{ab}\right)-\\ &\left(1-\theta\right)N_a\left(\varsigma_a\left(\gamma_k^{c,ab},y_k\right)+\overline{\gamma}_a^{c,ab}\overline{Y}_a\right)-\theta N_b\left(\varsigma_b\left(\gamma_k^{c,ab},y_k\right)+\overline{\gamma}_b^{c,ab}\overline{Y}_b\right)\end{aligned} \tag{4.7}$$

where $\overline{Y}_{ab}=\sum_{U_{ab}}y_k\Big/N_{ab}$, $\overline{\gamma}^{ab,c}=\sum_{U_{ab}}\gamma_k^{ab,c}\Big/N_{ab}$, $\overline{Y}_a=\sum_{U_a}y_k\Big/N_a$, $\overline{\gamma}_a^{c,ab}=\sum_{U_a}\gamma_k^{c,ab}\Big/N_a$, $\overline{Y}_b=\sum_{U_b}y_k\Big/N_b$ and $\overline{\gamma}_b^{c,ab}=\sum_{U_b}\gamma_k^{c,ab}\Big/N_b$. $\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)$ is the population covariance between the misclassification probabilities $\gamma_k^{ab,c}$ and the values of the target variable y_k within the overlapping domains *ab*. Also, $\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)$ are the population covariance between the misclassification probabilities $\gamma_k^{c,ab}$ and the values of the target variable y_k within the non-overlapping domains *a* and *b*, respectively. These covariances can be written as follows

$$\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right) = \sum\nolimits_{U_{ab}}\left(\gamma_k^{ab,c} - \bar{\gamma}^{ab,c}\right)\left(y_k - \bar{Y}_{ab}\right)\Big/N_{ab} \tag{4.8}$$

$$\varsigma_{a}\left(\gamma_k^{c,ab}, y_k\right) = \sum\nolimits_{U_{a}}\left(\gamma_k^{c,ab} - \bar{\gamma}_a^{c,ab}\right)\left(y_k - \bar{Y}_{a}\right)\Big/N_{a} \tag{4.9}$$

$$\varsigma_{b}\left(\gamma_k^{c,ab}, y_k\right) = \sum\nolimits_{U_{b}}\left(\gamma_k^{c,ab} - \bar{\gamma}_b^{c,ab}\right)\left(y_k - \bar{Y}_{b}\right)\Big/N_{b} \tag{4.10}$$

In the previous propositions, the bias in the dual frame estimators resulting from the misclassification problem was explored. Obviously, the misclassification bias depends on two components:

a) The expected total of y_k for the misclassified cases within each domain, $N_{ab}\bar{\gamma}^{ab,c}\bar{Y}_{ab}$, $N_a\bar{\gamma}_a^{c,ab}\bar{Y}_a$ and $N_b\bar{\gamma}_b^{c,ab}\bar{Y}_b$.

b) The correlation between the misclassifications probabilities and the study variable *y* within the different design domains, supported by the within domains covariances, $\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right)$, $\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)$.

This means that the misclassification bias can be controlled during the data collection process by following the best practices that decrease the measurement error in reporting the domain membership variable. At the same time, the misclassification bias can be adjusted based on the second component by implicitly predicting the misclassification probabilities. This can be performed by calibrating the data by an auxiliary variable that is correlated with the study variable y and the misclassification probabilities. This step can be performed either in the standard dual frame estimators or in the JCE. In the standard dual frame estimators, the calibration step comes after combining the data based on the misclassified domains. When misclassification probabilities are known, Lohr (2011) proposed an adjustment factor for the misclassification bias for the FWE estimator, which is consistent with our derivations of the misclassification bias.

In JCE, the domain misclassification does not affect the estimates as long as no domain membership information was added to the auxiliary variable vector, $\mathbf{x}$. However, even if misclassified domain membership information was added to the auxiliary variable vector, adding more auxiliary variables which are correlated with the study variable y and the misclassification probabilities is enough to reduce the bias resulted from the misclassified domain. Moreover, the effect of using the misclassified domains as the sole auxiliary variable in JCE is less significant than the effect of the domain misclassification in the standard dual frame estimators. This is due the fact that in the standard dual frame estimators, classifying the sampling units into the domain correctly is required before applying the composite factor θ. However, in JCE, this misclassification error is accounted for as a measurement in the auxiliary variables.

4.4 Simulation study

Simulation studies were used to evaluate the performance of the JCE relative to the standard dual frame estimators in the presence of different misclassification and nonresponse mechanisms. The same populations generated in Chapter 2 were used in this chapter. The simulation factors are as the following:

1. Sampling Designs: Simple Sampling Design where simple random samples were selected from both frames.

2. Sample size: Equal allocation where $n_A = 500$ and $n_B = 500$ under full response assumption and $n_A = 1500$ and $n_B = 1500$ in the presence of nonresponse.

3. Domain means: Large-differences domains' means where $\beta_a = 5,\ \beta_{ab} = 10$ and $\beta_b = 15$.

4. Correlation between y_{jk} and x_{jk} : The population correlation coefficient is $\rho_{xy} = 0.40$.

5. Response mechanisms

   a) Full Response Mechanism (FRM), where overall response rate is 100%.

   b) The same 6 response mechanisms in Chapter 3

      I. Simple Response Propensity Model (MCAR).

      II. Response Propensity by Auxiliary Variable *x* (MAR2).

      III. Response Propensity by Frame (MAR3).

      IV. Response Propensity by Frame and Auxiliary Variable x (MAR4).

      V. Response Propensity by Design Domain (MAR5).

      VI. Response Propensity by Design Domain and Auxiliary Variable *x* (MAR6).

6. Misclassification mechanisms

   a) The one-way OWOM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(ab,a)} = .1$ and $\gamma^{B(ab,b)} = .1$. This means that 10 % of the sample A overlapping domain *ab* cases are misclassified in non-overlapping domain *a* and 10 % of the sample B overlapping domain *ab* cases are misclassified in non-overlapping domain *b*.

   b) The one-way OWNM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = .1$ and $\gamma^{B(b,ab)} = .1$. This means that 10 % of the sample A non-overlapping domain *a* cases are misclassified in overlapping domain *ab* and 10 % of the sample B non-overlapping domain *b* cases are misclassified in overlapping domain *ab*.

   c) The two-way TWM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = .1$, $\gamma^{B(b,ab)} = .1$, $\gamma^{A(a,ab)} = .1$ and $\gamma^{B(b,ab)} = .1$.

These sets of simulation factors combine to form 42 simulation studies, 21 simulation studies for each population model. One thousand replicates of initial samples of 3,000 cases were run for each study where nonresponse was present. For the FRM response mechanism, the initial sample sizes were 1,000 cases. To simulate a dual frame design, within each simulation replicate, two equal-size samples were drawn separately from both frames A and B, where $n_A = n_B = 1,500$, and $n_A = n_B = 500$ for the FRM response mechanism. These samples were 'stacked' to form dual frame sample *s*. Conditional on the misclassification and response mechanisms, the misclassified response sets r_A and r_B were realized and the misclassified domains were generated.

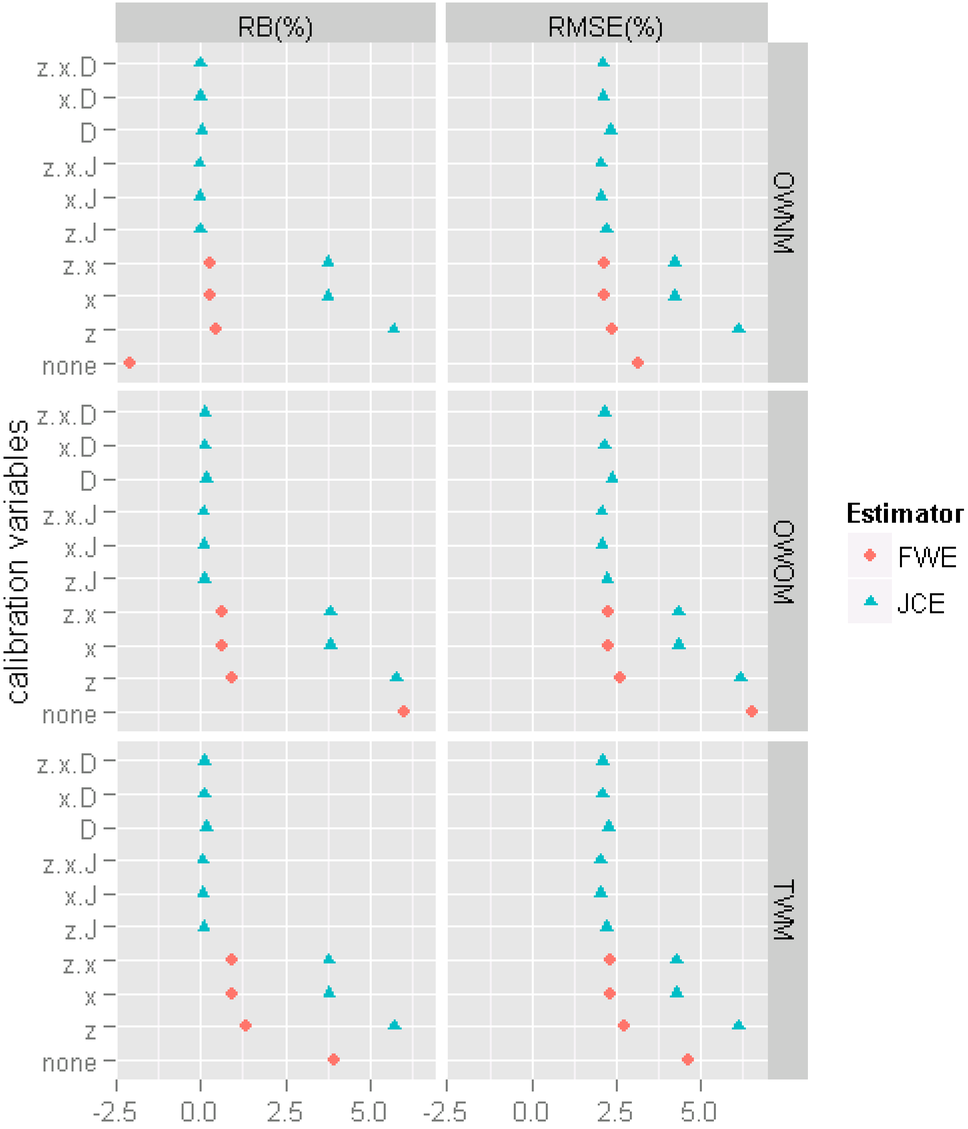
4.5 Simulation Results

Generally, as indicated in Tables 4.1 and 4.3, in the presence of non-sampling errors, domain misclassification or nonresponse, biases in $\hat{Y}_{FWE}$ are present. Even under the complete response, in Figure 4.1, domain misclassification results in biased $\hat{Y}_{FWE}$. The bias magnitude varies based on the misclassification mechanism; the one-way OWNM and the two-way TWM mechanisms result in smaller relative biases than the one-way OWOM does, as in Figure 4.1. Under the complete response assumption, the biases that result under the OWOM mechanism are positive sign biases, however it is negative sign biases under the OWNM. The summation of the biases resulting from the OWOM and OWNM mechanisms equals the net bias resulting from the TWM mechanism. Since the effect of adding the nonresponse besides the misclassification is the same as its effect as a sole non-sampling error, discussed in Chapter 3, we will only highlight the effect of the misclassification error.

Under the CLR model, in Table 4.1, the standard estimator $\hat{Y}_{FWE}$ is affected by the misclassification error, whereas the proposed estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are not. Adding the calibration in the standard estimators $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$ reduces the misclassification bias and achieved relative biases comparable to the JCE estimators, $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Interestingly, adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. Even calibrating only by the misclassified domains in $\hat{Y}_{JCE.D}$ results in almost unbiased estimates. Generally, the relative mean square errors show same patterns as the relative biases, as indicated in Table 4.2. However, MSE's for $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.x}$ were slightly lower than MSE's for $\hat{Y}^{cal}_{FWE.z}$ and $\hat{Y}^{cal}_{FWE.x}$, respectively.

Under the GLR model, in Figure 4.1 and Table 4.3, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. However, the relative biases in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ were smaller than the standard estimator $\hat{Y}_{FWE}$. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zJ}$, $\hat{Y}_{JCE.xJ}$ and $\hat{Y}_{JCE.xzJ}$ resulted in reduced relative biases. Clearly in Figure 4.1, adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.D}$, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. The relative mean square errors show similar patterns to relative biases, as indicated in Figure 4.1 and Table 4.4. However, MSE's for $\hat{Y}_{JCE.zL}$, $\hat{Y}_{JCE.xL}$ and $\hat{Y}_{JCE.xzL}$ were slightly lower than MSE's for $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively.

Figure 4.1: Simulation RB (%) and RMSE (%) for FWE and JCE estimators estimated from the GLR model population under full response and $\rho_{xy} = 0.40$



4.6 Discussion and Conclusion

In this chapter, the domain misclassification was introduced as a form of the non-sampling error, which could affect the bias properties of the dual frame estimators. The effect of the domain misclassification exceeds its effect as a type of measurement or reporting error in the domain membership information. The misclassified domains may affect the standard dual frame estimators substantially. This is due to the fact that the standard dual frame estimators require accurate information about the domain membership. Based on this information, the adjustment factor is applied to the design weights for dual frame estimation.

In this chapter, we derived expressions for the analytic bias that results when the standard dual frame estimators are applied to data with different domain misclassification mechanisms. These bias expressions indicated that the correlation between the misclassification probabilities and the study variable y within each domain is an important determinant of the misclassification bias. Also, the expected total of the y variable for the misclassified cases within each domain is another determinant of the misclassification bias. Controlling these two determinants could be the key for reducing the misclassification bias in the standard dual frame estimators.

The misclassification bias can be reduced by following the best practices during the data collection, under which the measurement errors are controlled especially for the domain membership variable. Also, calibrating the data using the auxiliary variables which are correlated with the study variable and the misclassification probabilities could be a promising approach for adjusting the misclassification error. This approach is motivated by the domain-level correlation between the misclassification probabilities and the study variable y as a determinant for the misclassification bias. In the simulation studies, calibrating the standard dual frame estimators with x, which is correlated with y, or with z, which is correlated with the misclassification probabilities, due to the high correlation with the domains, was enough to adjust for the domain misclassification bias.

In addition to introducing the domain misclassification problem in this chapter, the JCE was introduced as a robust dual frame estimator to the domain misclassification error. The JCE does not necessarily need any information about the domain classification. Therefore, the misclassification problem does not affect the JCE estimates as long as the domain membership information was not added to the calibration auxiliary variable vector. Interestingly, adding the misclassified domains to the JCE auxiliary variable vector does not lead to substantially biased estimates, as long as the domains are misclassified at random. This is due to the fact that the effect of the misclassified domains in the context of the JCE is a measurement error effect. Moreover, under the GLM model, calibrating the dual frame samples in JCE by the misclassified domains ignoring the strata totals was enough to result in reduced bias estimates. This is due to the effect of adding the domain membership in the calibration auxiliary variable vector. As discussed in (2.28) and (2.29) in Chapter 2, adding this information results in unbiased estimates. However, the measurement error in the domain membership results in slightly biased estimates.

Table 4.1: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under $\rho_{xy} = 0.40$ and in the presence of the misclassification problem.

Non-response	Misclassification	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
FRM	OWOM	5.02	-0.04	-0.07	0.00	-0.03	0.00	-0.03	-0.05	-0.02	-0.02
MCAR	OWOM	-68.48	0.05	-0.02	0.01	-0.03	0.01	-0.03	0.05	0.01	0.01
MAR2	OWOM	-67.41	3.55	3.53	0.07	0.06	0.07	0.06	3.54	0.05	0.05
MAR3	OWOM	-68.10	0.13	0.04	0.08	0.03	0.08	0.03	0.12	0.08	0.08
MAR4	OWOM	-66.87	3.69	3.64	0.09	0.08	0.09	0.08	3.67	0.07	0.07
MAR5	OWOM	-70.31	-0.05	-0.12	-0.06	-0.11	-0.05	-0.11	0.01	0.00	0.00
MAR6	OWOM	-68.92	3.37	3.12	0.03	0.00	0.03	0.00	3.57	0.06	0.06
FRM	OWNM	-2.46	0.02	-0.05	0.03	-0.03	0.02	-0.03	0.03	0.03	0.02
MCAR	OWNM	-70.77	-0.06	-0.12	-0.02	-0.06	-0.02	-0.06	-0.05	-0.02	-0.02
MAR2	OWNM	-69.79	3.54	3.49	-0.05	-0.07	-0.05	-0.08	3.55	-0.05	-0.05
MAR3	OWNM	-70.21	0.07	0.01	-0.01	-0.02	-0.01	-0.02	0.03	-0.02	-0.02
MAR4	OWNM	-69.21	3.66	3.64	0.01	0.02	0.01	0.02	3.63	-0.01	0.00
MAR5	OWNM	-72.39	-0.04	-0.12	-0.02	-0.05	-0.02	-0.05	0.00	0.01	0.01
MAR6	OWNM	-71.29	3.34	3.04	0.01	0.00	0.01	0.00	3.56	0.02	0.02
FRM	TWM	2.48	-0.12	-0.12	-0.10	-0.07	-0.10	-0.07	-0.09	-0.08	-0.07
MCAR	TWM	-69.18	0.07	0.02	0.06	0.03	0.06	0.03	0.10	0.06	0.06
MAR2	TWM	-67.91	3.59	3.57	0.06	0.05	0.06	0.05	3.57	0.04	0.04
MAR3	TWM	-68.60	0.05	-0.02	0.04	-0.02	0.04	-0.02	0.03	0.02	0.02
MAR4	TWM	-67.73	3.68	3.64	0.04	0.04	0.04	0.04	3.67	0.04	0.04
MAR5	TWM	-70.56	0.10	0.03	0.05	0.02	0.04	0.02	0.13	0.09	0.09
MAR6	TWM	-69.62	3.30	3.11	0.02	0.01	0.02	0.01	3.49	0.02	0.03

Table 4.2: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the CLR model population under $\rho_{xy} = 0.40$ and in the presence of the misclassification problem.

Non-response	Misclassification	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
FRM	OWOM	5.55	1.94	1.84	1.78	1.69	1.78	1.70	1.94	1.78	1.78
MCAR	OWOM	68.49	2.07	1.95	1.89	1.79	1.89	1.79	2.08	1.91	1.91
MAR2	OWOM	67.42	4.10	3.99	1.91	1.74	1.91	1.74	4.10	1.93	1.93
MAR3	OWOM	68.11	2.02	1.91	1.89	1.75	1.88	1.75	1.99	1.86	1.86
MAR4	OWOM	66.88	4.22	4.13	1.91	1.79	1.91	1.79	4.18	1.88	1.88
MAR5	OWOM	70.32	2.15	2.00	1.93	1.83	1.93	1.83	2.42	2.16	2.16
MAR6	OWOM	68.93	3.97	3.69	1.97	1.83	1.97	1.84	4.21	2.11	2.12
FRM	OWNM	3.31	1.93	1.79	1.77	1.63	1.77	1.63	1.94	1.79	1.79
MCAR	OWNM	70.78	2.07	1.89	1.88	1.72	1.89	1.72	2.07	1.89	1.89
MAR2	OWNM	69.79	4.07	3.96	1.86	1.75	1.86	1.75	4.08	1.87	1.87
MAR3	OWNM	70.22	2.15	1.97	1.96	1.80	1.96	1.80	2.09	1.90	1.91
MAR4	OWNM	69.22	4.23	4.15	1.98	1.87	1.99	1.87	4.18	1.95	1.95
MAR5	OWNM	72.40	2.05	1.94	1.87	1.77	1.88	1.78	2.28	2.06	2.07
MAR6	OWNM	71.30	3.91	3.59	1.91	1.81	1.91	1.81	4.18	2.05	2.06
FRM	TWM	3.43	1.92	1.82	1.74	1.64	1.74	1.64	1.90	1.72	1.72
MCAR	TWM	69.19	2.08	1.93	1.93	1.79	1.92	1.79	2.11	1.95	1.94
MAR2	TWM	67.92	4.10	4.04	1.88	1.78	1.89	1.78	4.09	1.88	1.89
MAR3	TWM	68.61	2.14	1.99	1.96	1.82	1.96	1.82	2.14	1.96	1.96
MAR4	TWM	67.74	4.24	4.16	1.96	1.84	1.96	1.85	4.22	1.93	1.93
MAR5	TWM	70.57	2.15	2.01	1.98	1.84	1.98	1.84	2.33	2.16	2.16
MAR6	TWM	69.63	3.91	3.68	1.93	1.79	1.93	1.79	4.16	2.05	2.05

Table 4.3: Simulation RB (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under $\rho_{xy} = 0.40$ and in the presence of the misclassification problem.

Non-response	Mis-classification	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
FRM	OWOM	6.05	0.95	5.78	0.63	3.84	0.63	3.83	0.10	0.07	0.07	0.16	0.12	0.12
MCAR	OWOM	-68.28	0.69	5.61	0.45	3.70	0.46	3.70	-0.10	-0.10	-0.10	-0.12	-0.10	-0.10
MAR2	OWOM	-66.89	5.81	10.44	0.67	3.92	0.67	3.92	4.23	0.09	0.09	4.49	0.13	0.12
MAR3	OWOM	-67.96	0.59	5.43	0.40	3.58	0.41	3.59	0.08	0.03	0.04	0.11	0.08	0.08
MAR4	OWOM	-66.54	5.51	10.10	0.37	3.59	0.38	3.60	4.20	-0.03	-0.03	4.48	0.02	0.02
MAR5	OWOM	-68.77	4.83	9.01	3.20	6.02	3.20	6.02	-0.09	-0.07	-0.07	-0.05	-0.04	-0.04
MAR6	OWOM	-67.45	8.99	12.73	3.02	5.92	3.03	5.92	4.14	-0.09	-0.09	4.51	-0.05	-0.05
FRM	OWNM	-2.08	0.43	5.71	0.27	3.75	0.27	3.75	-0.02	-0.05	-0.05	0.01	-0.01	-0.01
MCAR	OWNM	-70.64	0.40	5.62	0.26	3.70	0.26	3.70	-0.06	-0.07	-0.07	-0.05	-0.04	-0.04
MAR2	OWNM	-69.59	5.31	10.33	0.25	3.78	0.26	3.79	4.07	-0.08	-0.08	4.34	-0.04	-0.04
MAR3	OWNM	-70.30	0.12	5.28	0.09	3.49	0.09	3.49	-0.07	-0.07	-0.07	-0.04	-0.03	-0.03
MAR4	OWNM	-69.20	5.28	10.18	0.14	3.59	0.14	3.59	4.26	-0.05	-0.05	4.56	-0.01	0.00
MAR5	OWNM	-71.36	4.60	9.08	3.06	6.06	3.06	6.06	0.04	0.03	0.03	0.08	0.06	0.07
MAR6	OWNM	-69.88	8.57	12.53	2.70	5.78	2.71	5.78	4.21	-0.08	-0.08	4.62	-0.01	-0.01
FRM	TWM	3.96	1.34	5.74	0.91	3.80	0.91	3.80	0.07	0.06	0.06	0.14	0.11	0.11
MCAR	TWM	-68.91	0.98	5.46	0.63	3.58	0.63	3.58	-0.30	-0.25	-0.25	-0.25	-0.20	-0.21
MAR2	TWM	-67.54	6.16	10.38	0.80	3.77	0.80	3.77	4.16	-0.04	-0.04	4.38	-0.04	-0.04
MAR3	TWM	-68.59	0.87	5.29	0.60	3.51	0.60	3.51	-0.04	-0.03	-0.04	-0.01	0.00	0.00
MAR4	TWM	-67.17	5.99	10.17	0.59	3.54	0.60	3.54	4.30	-0.05	-0.05	4.55	-0.04	-0.03
MAR5	TWM	-69.17	5.19	8.94	3.44	5.94	3.44	5.94	-0.08	-0.09	-0.08	-0.04	-0.02	-0.01
MAR6	TWM	-68.11	9.38	12.73	3.28	5.87	3.29	5.87	4.24	-0.02	-0.02	4.64	0.03	0.03

Table 4.4: Simulation RMSE (%) for FWE and JCE estimators of $\hat{Y}$, estimated from the GLR model population under $\rho_{xy} = 0.40$ and in the presence of the misclassification problem.

Non-response	Mis-classification	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zJ}$	$\hat{Y}_{JCE.xJ}$	$\hat{Y}_{JCE.xzJ}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
FRM	OWOM	6.58	2.63	6.20	2.30	4.35	2.30	4.35	2.25	2.08	2.08	2.38	2.18	2.17
MCAR	OWOM	68.29	2.61	6.07	2.28	4.25	2.28	4.25	2.34	2.12	2.12	2.44	2.20	2.20
MAR2	OWOM	66.90	6.38	10.71	2.57	4.56	2.57	4.56	4.93	2.38	2.39	5.17	2.42	2.42
MAR3	OWOM	67.97	2.72	5.99	2.38	4.23	2.38	4.23	2.47	2.25	2.25	2.54	2.28	2.28
MAR4	OWOM	66.56	6.06	10.37	2.42	4.23	2.42	4.24	4.85	2.31	2.31	5.11	2.32	2.32
MAR5	OWOM	68.78	5.48	9.31	3.95	6.39	3.96	6.39	2.67	2.44	2.45	2.76	2.51	2.51
MAR6	OWOM	67.46	9.34	12.94	3.86	6.32	3.86	6.32	4.94	2.54	2.54	5.31	2.62	2.62
FRM	OWNM	3.15	2.43	6.13	2.19	4.25	2.19	4.25	2.24	2.07	2.07	2.33	2.13	2.13
MCAR	OWNM	70.64	2.52	6.08	2.29	4.27	2.30	4.27	2.37	2.21	2.21	2.43	2.24	2.25
MAR2	OWNM	69.60	5.90	10.60	2.38	4.39	2.38	4.39	4.74	2.27	2.27	5.01	2.33	2.33
MAR3	OWNM	70.31	2.59	5.79	2.30	4.09	2.30	4.09	2.43	2.19	2.20	2.47	2.22	2.22
MAR4	OWNM	69.21	5.90	10.46	2.43	4.26	2.43	4.26	4.92	2.29	2.29	5.22	2.35	2.35
MAR5	OWNM	71.37	5.28	9.40	3.84	6.44	3.85	6.44	2.75	2.46	2.47	2.83	2.52	2.52
MAR6	OWNM	69.89	8.94	12.75	3.62	6.20	3.62	6.20	4.96	2.44	2.44	5.34	2.49	2.49
FRM	TWM	4.68	2.75	6.16	2.34	4.31	2.35	4.31	2.24	2.06	2.06	2.30	2.10	2.10
MCAR	TWM	68.92	2.77	5.96	2.39	4.18	2.39	4.18	2.39	2.19	2.19	2.47	2.24	2.24
MAR2	TWM	67.55	6.66	10.64	2.51	4.37	2.51	4.37	4.82	2.27	2.27	5.06	2.37	2.37
MAR3	TWM	68.60	2.66	5.80	2.36	4.12	2.36	4.12	2.37	2.20	2.20	2.44	2.24	2.24
MAR4	TWM	67.19	6.54	10.45	2.50	4.19	2.51	4.20	4.94	2.30	2.30	5.20	2.36	2.36
MAR5	TWM	69.18	5.76	9.24	4.12	6.30	4.12	6.31	2.58	2.38	2.38	2.67	2.45	2.45
MAR6	TWM	68.12	9.72	12.95	4.07	6.28	4.07	6.28	4.99	2.44	2.44	5.36	2.50	2.50

Appendix 4.1: Proof of Proposition 4.1

$$\hat{Y}_{mis} = \sum\nolimits_{s_c} d_k y_k + \sum\nolimits_{s_{ab}} \left(1 - I_k^{ab,c}\right) m_k d_k y_k + \sum\nolimits_{s_{ab}} I_k^{ab,c} d_k y_k \tag{4.11}$$

$$\hat{Y}_{mis} = \sum\nolimits_{s_a} d_k y_k + \sum\nolimits_{s_b} d_k y_k + \sum\nolimits_{s_{ab}} m_k d_k y_k - \sum\nolimits_{s_{ab}} I_k^{ab,c} m_k d_k y_k + \sum\nolimits_{s_{ab}} I_k^{ab,c} d_k y_k \tag{4.12}$$

Note that $$E_{pq}\left(\delta_k . I_k^{ab,c}\right) = E_p\left(\delta_k\right).E_{q|p}\left(I_k^{ab,c}\right) = \pi_k . \gamma_k^{ab,c} \tag{4.13}$$

where $E_{pq}(\ )$ denote the joint expectation with respect to sampling distribution $p(s)$ and misclassification distribution $q(s_{mis} \mid s)$ where s_{mis} is the sample *s* with misclassification. Where δ_k is a sampling indicator for observation *k*, $E_p\left(\delta_k\right) = \pi_k$ which is the sample selection probability. Also, $E_{q|p}\left(I_k^{ab,c}\right) = \gamma_k^{ab,c}$ is the conditional probability of misclassification from the overlapping domains to non-overlapping domains.

$$E_{pq}\left[\hat{Y}_{mis}\right] = Y_a + Y_b + Y_{ab} + E_{pq}\left(\sum\nolimits_{s_{ab}} I_k^{ab,c} d_k y_k - \sum\nolimits_{s_{ab}} I_k^{ab,c} m_k d_k y_k\right) \tag{4.14}$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E_{pq}\left(\hat{Y}_{mis}\right) - Y = E_{pq}\left(\sum\nolimits_{s_{ab}} I_k^{ab,c} d_k y_k - \sum\nolimits_{s_{ab}} I_k^{ab,c} m_k d_k y_k\right) \tag{4.15}$$

$$= E_{pq}\left(\sum\nolimits_{s_{ab}^A} I_k^{ab,c} d_k y_k + \sum\nolimits_{s_{ab}^B} I_k^{ab,c} d_k y_k - \theta \sum\nolimits_{s_{ab}^A} I_k^{ab,c} d_k y_k - (1-\theta) \sum\nolimits_{s_{ab}^B} I_k^{ab,c} d_k y_k\right) \tag{4.16}$$

$$= E_{pq}\left(\sum\nolimits_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k + \sum\nolimits_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k - \theta \sum\nolimits_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k - (1-\theta) \sum\nolimits_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k\right) \tag{4.17}$$

$$= E_{pq}\left(\sum\nolimits_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k\right) = \sum\nolimits_{U_{ab}} E_p\left(\delta_k\right) E_{q/p}\left(I_k^{ab,c}\right) d_k y_k = \gamma^{ab,c} \sum\nolimits_{U_{ab}} y_k \tag{4.18}$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = \gamma^{ab,c} Y_{ab} \tag{4.19}$$

where $\gamma_k^{ab,c} = \gamma^{ab,c}$ for all *k*.

Appendix 4.2: Proof of Proposition 4.2

$$\hat{Y}_{mis} = \sum_{s_{ab}} m_k d_k y_k + \sum_{s_c} \left(1 - I_k^{c,ab}\right) d_k y_k + \sum_{s_c} I_k^{c,ab} m_k d_k y_k \tag{4.20}$$

$$\hat{Y}_{mis} = \sum_{s_a} d_k y_k + \sum_{s_b} d_k y_k + \sum_{s_{ab}} m_k d_k y_k - \sum_{s_c} I_k^{c,ab} d_k y_k + \sum_{s_c} I_k^{c,ab} m_k d_k y_k \tag{4.21}$$

$$E_{pq}\left(\hat{Y}_{mis}\right) = Y + E\left(\sum_{s_c} I_k^{c,ab} m_k d_k y_k - \sum_{s_c} I_k^{c,ab} d_k y_k\right) \tag{4.22}$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E\left(\sum_{s_c} I_k^{c,ab} m_k d_k y_k - \sum_{s_c} I_k^{c,ab} d_k y_k\right) \tag{4.23}$$

$$= E_{pq}\left(\sum_{s_a} I_k^{c,ab} m_k d_k y_k + \sum_{s_b} I_k^{c,ab} m_k d_k y_k - \sum_{s_a} I_k^{c,ab} d_k y_k - \sum_{s_b} I_k^{c,ab} d_k y_k\right) \tag{4.24}$$

$$= E_{pq}\left(\theta \sum_{s_a} I_k^{c,ab} d_k y_k + (1-\theta) \sum_{s_b} I_k^{c,ab} d_k y_k - \sum_{s_a} I_k^{c,ab} d_k y_k - \sum_{s_b} I_k^{c,ab} d_k y_k\right) \tag{4.25}$$

$$= E_{pq}\left((\theta - 1) \sum_{U_a} \delta_k I_k^{c,ab} d_k y_k - \theta \sum_{U_b} \delta_k I_k^{c,ab} d_k y_k\right) \tag{4.26}$$

$$= \left((\theta - 1) \gamma^{c,ab} \sum_{U_a} y_k - \theta \gamma^{c,ab} \sum_{U_b} y_k\right) \tag{4.27}$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = \gamma^{c,ab} \left((\theta - 1) \sum_{U_a} y_k - \theta \sum_{U_b} y_k\right) \tag{4.28}$$

Where $E_{q|p}\left(I_k^{c,ab}\right) = \gamma^{c,ab}$ which is the probability of misclassification from the non-overlapping domains to overlapping domains.

Appendix 4.3: Proof of Proposition 4.4

Under the two-way TWM misclassification

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=E_{pq}\left(\sum\nolimits_{U_{ab}}\delta_k I_k^{ab,c}d_k y_k\right)+E_{pq}\left(\left(\theta-1\right)\sum\nolimits_{U_a}\delta_k I_k^{c,ab}d_k y_k-\theta\sum\nolimits_{U_b}\delta_k I_k^{c,ab}d_k y_k\right) \quad (4.29)$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right)=\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k+\left(\theta-1\right)\sum\nolimits_{U_a}\gamma_k^{c,ab}y_k-\theta\sum\nolimits_{U_b}\gamma_k^{c,ab}y_k \quad (4.30)$$

$$\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k=\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k+\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}-\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c} \quad (4.31)$$

$$=\left(N_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}-N_{ab}\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}\right)\Big/N_{ab} \quad (4.32)$$

$$=\left(N_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}\right)\Big/N_{ab}-N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c} \quad (4.33)$$

$$=N_{ab}\left(\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}y_k+N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c}-\bar{Y}_{ab}\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}+\bar{\gamma}^{ab,c}\sum\nolimits_{U_{ab}}y_k\right)\Big/N_{ab}-N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c} \quad (4.34)$$

$$=N_{ab}\sum\nolimits_{U_{ab}}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\Big/N_{ab}+N_{ab}\bar{Y}_{ab}\bar{\gamma}^{ab,c} \quad (4.35)$$

$$=N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)+\bar{Y}_{ab}\bar{\gamma}^{ab,c}\right) \quad (4.36)$$

where

$$\varsigma_{ab}\left(\gamma_k^{ab,c},y_k\right)=\sum\nolimits_{U_{ab}}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\Big/N_{ab} \quad (4.37)$$

where $\bar{Y}_{ab}=\sum\nolimits_{U_{ab}}y_k\Big/N_{ab}$ and $\bar{\gamma}^{ab,c}=\sum\nolimits_{U_{ab}}\gamma_k^{ab,c}\Big/N_{ab}$.

Similarly

$$\sum\nolimits_{U_a}\gamma_k^{c,ab}y_k=N_a\left(\varsigma_a\left(\gamma_k^{c,ab},y_k\right)+\bar{\gamma}_a^{c,ab}\bar{Y}_a\right) \quad (4.38)$$

and

$$\sum\nolimits_{U_b} \gamma_k^{c,ab} y_k = N_b \left(\varsigma_b \left(\gamma_k^{c,ab}, y_k \right) + \bar{\gamma}_b^{c,ab} \bar{Y}_b \right) \tag{4.39}$$

where

$$\varsigma_a \left(\gamma_k^{c,ab}, y_k \right) = \sum\nolimits_{U_a} \left(\gamma_k^{c,ab} - \bar{\gamma}_a^{c,ab} \right) \left(y_k - \bar{Y}_a \right) \Big/ N_a \tag{4.40}$$

$$\varsigma_b \left(\gamma_k^{c,ab}, y_k \right) = \sum\nolimits_{U_b} \left(\gamma_k^{c,ab} - \bar{\gamma}_b^{c,ab} \right) \left(y_k - \bar{Y}_b \right) \Big/ N_b \tag{4.41}$$

where $\bar{Y}_a = \sum\nolimits_{U_a} y_k \Big/ N_a$, $\bar{Y}_b = \sum\nolimits_{U_b} y_k \Big/ N_b$, $\bar{\gamma}_a^{c,ab} = \sum\nolimits_{U_a} \gamma_k^{c,ab} \Big/ N_a$ and $\bar{\gamma}_b^{c,ab} = \sum\nolimits_{U_b} \gamma_k^{c,ab} \Big/ N_b$.

$$\begin{gathered} Bias_{pq} \left(\hat{Y}_{mis} \right) = N_{ab} \left(\varsigma_{ab} \left(\gamma_k^{ab,c}, y_k \right) + \bar{\gamma}^{ab,c} \bar{Y}_{ab} \right) - \\ \left(1 - \theta \right) N_a \left(\varsigma_a \left(\gamma_k^{c,ab}, y_k \right) + \bar{\gamma}_a^{c,ab} \bar{Y}_a \right) - \theta N_b \left(\varsigma_b \left(\gamma_k^{c,ab}, y_k \right) + \bar{\gamma}_b^{c,ab} \bar{Y}_b \right) \end{gathered} \tag{4.42}$$

Chapter 5

Joint Calibration as a Model-based Approach for Dual Frame Estimation

5.1 Introduction

In chapter 1, the dual frame estimation problem was introduced in the context of *the probability sampling theory*. In chapters 2, 3 and 4, the JCE was introduced as a model-assisted design-based approach for dual frame estimation. In this chapter, we will explore the dual frame estimation problem in the context of *the prediction theory*. The correspondence between the model-based dual frame estimation and the design-based joint calibration will be explored. A preface for the prediction theory is introduced in Section 5.2. The dual frame estimation problem is discussed in the context of the prediction theory in Section 5.3. Different model-based estimators are derived and compared with the JCE estimators in Section 5.4. A conclusion is discussed in Section 5.5.

5.2 The Prediction Theory

In this section, we briefly discuss the difference between probability sampling theory and prediction theory as a basis for design-based and model-based estimations, respectively. The following example shows the difference between the two approaches by presenting the well-known ratio estimator based on the sampling and prediction theories.

Example 5.2.1

Suppose that we have data collected on two continuous variables *y* and *x* in sample *s* from a population *U* of size *N*, where *y* is known only for the sampled cases in *s*, assume *x* is known for all cases in *U*, which means that $X = \sum_U x_k$ is known. We want to estimate the population total of *y*, $Y = \sum_U y_k$.

Under the probability sampling theory, selection probabilities $\pi_k = p(k \in s)$ are used to adjust the sampled cases in *s*. The estimated total of *y* can be written as

$$\hat{Y} = \sum_s \pi_k^{-1} y_k \tag{5.1}$$

where π_k^{-1} work as the design weight that adjusts for the sampling selection design. This estimator in (5.1) is a design-unbiased estimator for the population total *Y* as follows

$$E\left(\hat{Y}\right) = E\left(\sum_s \pi_k^{-1} y_k\right) = E\left(\sum_U I_k \pi_k^{-1} y_k\right) = Y \tag{5.2}$$

where *E* denotes the expectation with regard to the sampling selection, I_k is a sample selection identifier variable which equal 1 for the sampled cases in *s* and 0 for the non-sampled cases in *c* and $E\left(I_k\right) = \pi_k$. The estimator in (5.1) is the well-known Horvitz-Thompson estimator proposed by Horvitz and Thompson (1952). Under simple random sampling design without replacement (srswor), the estimator in (5.1) can be written as

$$\hat{Y} = N\overline{y}_s \tag{5.3}$$

where $\bar{y}_s = \sum_s y_k / n$.

When population totals *N* and *X* are known, under srswor, the well-known ratio estimator can be written as

$$\hat{Y} = N\bar{X}\,\bar{y}_s / \bar{x}_s \tag{5.4}$$

where $\bar{x}_s = \sum_s x_k / n$ and $\bar{X} = \sum_U x_k / N$.

Under prediction theory, the population total of y can be written as

$$Y = \sum_s y_k + \sum_c y_k \tag{5.5}$$

where *c* is a set of the non-sampled cases. In these settings, predicting *y* for the non-sampled cases and consequently predicting $\sum_c y_k$ to estimate $\hat{Y}$ is the main idea behind the prediction theory (Valliant, Dorfman, & Royall, 2000). This can be performed by modeling the sampled cases in *s* and then the fitted model can be used to predict $\sum_c y_k$ by $\beta \sum_c x_k$. Where the estimated total can be written as

$$\hat{Y} = \sum_s y_k + \beta \sum_c x_k = \sum_s y_k + \left[\frac{\hat{Y} - \sum_s y_k}{\sum_c x_k} \right] \sum_c x_k \tag{5.6}$$

$\left[\frac{\hat{Y}-\sum_s y_k}{\sum_c x_k}\right]$ works as an implicit estimator for β. Generally, the Best Linear Unbiased (BLU) estimator of Y can be achieved by using the BLU estimator $\hat{\beta}$. Under the ratio model, where the expectation and variance of y_k are

$$\begin{cases} E_{\xi}\left(y_k\right)=\beta x_k \\ V_{\xi}\left(y_k\right)=\sigma^2 x_k \end{cases} \tag{5.7}$$

following the general prediction theorem, the BLU estimator $\hat{\beta}$ can be written as

$$\hat{\beta}=\overline{y}_s/\overline{x}_s \tag{5.8}$$

Therefore, substituting β in (5.6) by $\hat{\beta}$ in (5.8) results in the well-known ratio estimator as in (5.4)

$$\hat{Y}=\sum_s y_k+\hat{\beta}\sum_c x_k=N\overline{X}\,\overline{y}_s/\overline{x}_s \tag{5.9}$$

In this example we highlighted the fact that the results of both the probability sampling theory and the prediction theory may coincide. This happens when the model-based BLU estimators reduce to familiar design-based estimators. The same property will be explored in the next two sections for the dual frame estimation problem.

5.3 The Dual Frame Estimation Problem Under The Prediction Theory

In this section, the dual frame estimation problem will be discussed in the context of the prediction theory. Since the dual frame estimation problem is a multiplicity problem resulting from the fact that some cases have more than one chance of being selected in the survey as discussed in (Mecatti, 2007), the same problem should have different properties under the prediction theory. This is because the prediction theory depends less heavily on the probability sampling design and depend more on the relationship between the variables (Valliant, Dorfman, & Royall, 2000). Under the prediction theory, the population total of y based on a dual frame design from population $U = U_A \cup U_B$ can be written as

$$Y = \sum_{s_A} y_k + \sum_{s_B} y_k - \sum_{s_d} y_k + \sum_{c} y_k \tag{5.10}$$

where $s_d = s_A \cap s_B$, a subset of duplicates, and $c = c_A \cup c_B$, c_A and c_B are the non-sampled cases from frame A and B, respectively. Where $\sum_{s_d} y_k$ is known, the dual frame estimation problem is to predict $\sum_{c} y_k$ after excluding the duplicates, $k \in s_d$. A weighted version of (5.10) can be written as

$$Y = \sum_{s} w_k^d y_k - \sum_{s} \delta_k^d y_k + \sum_{c} y_k \tag{5.11}$$

where $s = s_A \cup s_B$, w_k^d is a weighting variable to account for the duplicates and δ_k^d is an identifier variable for duplicates. w_k^d and δ_k^d can be defined as

$$w_k^d = \begin{cases} 2 & k \in s_A \cap s_B \\ 1 & k \notin s_A \cap s_B \end{cases} \quad \text{and} \quad \delta_k^d = \begin{cases} 1 & k \in s_A \cap s_B \\ 0 & k \notin s_A \cap s_B \end{cases} \tag{5.12}$$

Note that $\sum_{s} w_k^d y_k - \sum_{s} \delta_k^d y_k = \sum_{s} y_k$. This means that as long as the duplicates are identifiable and $\sum_{s} \delta_k^d y_k$ can be subtracted from $\sum_{s} w_k^d y_k$, the dual frame estimation problem becomes identical to a single frame estimation problem, where the main interest is to predict $\sum_c y_k$.

Therefore, where the estimated total $\hat{Y}$ can be written as

$$\hat{Y} = \sum_s y_k + \left[\frac{\hat{Y} - \sum_s y_k}{\sum_c x_k}\right] \sum_c x_k \tag{5.13}$$

we need to find an unbiased estimator for the parameter $\hat{\beta} = \left[\frac{\hat{Y} - \sum_s y_k}{\sum_c x_k}\right]$, where $\hat{\beta}$ works as an implicit estimator of the population model parameter β . Where the estimation error can be written as

$$\hat{Y} - Y = \hat{\beta} \sum_c x_i - \beta \sum_c x_i \tag{5.14}$$

$\hat{Y}$ is model-unbiased if $\left[E_\xi\left(\hat{\beta}\right) - \beta\right] \sum_c x_i = 0$, where the subscript ξ denotes the expectation with respect to the prediction model. With regard to the estimation error variance of an estimator $\hat{Y}$, it can be derived as

$$v_\xi\left(\hat{Y} - Y\right) = \left(\sum_c x_i\right)^2 v_\xi\left(\hat{\beta}\right) + v_\xi\left(\sum_c y_i\right) \tag{5.15}$$

In order to obtain the BLU estimator $\hat{Y}$, we need to minimize the error variance in (5.15), which requires the BLU estimator $\hat{\beta}$. This means that, under the prediction theory, the dual frame estimation problem reduces to a single frame estimation problem and so the general prediction theorem applies.

5.4 Model-based Dual Frame Estimation Examples

In this section, different model-based dual frame estimators are derived under different population models. These estimators will be compared with the JCE estimators derived in Chapter 2. Since all the derivations follow the general prediction theorem for the single frame prediction problem, we will not present the derivations. More details on the general prediction theorem can be found in (Valliant, Dorfman, & Royall, 2000).

Example 5.4.1: ***Common Ratio Model and Ratio Estimator***

Under the ratio model in (5.7), the same results in example 5.2.1 apply under the dual frame estimation. The ratio estimator in (5.9) is identical to the JCE which can be written as

$$\hat{Y}_{JCE} = N\bar{X}\left(\sum\nolimits_{s_A} y_k + \sum\nolimits_{s_B} y_k\right) \Big/ \left(\sum\nolimits_{s_A} x_k + \sum\nolimits_{s_B} x_k\right) \tag{5.16}$$

When the probability of sampling duplicates, $k \in s_A \cap s_B$, is ignorable.

Example 5.4.2: ***Common Mean Model and Expansion Estimator***

Under the common mean model, where the expectation and variance of y_k are

$$\begin{cases} E_\xi\left(y_k\right) = \mu \\ V_\xi\left(y_k\right) = \sigma^2 \end{cases} \tag{5.17}$$

the BLU estimator $\hat{\beta}$ can be written as

$$\hat{\beta} = \bar{y}_s \tag{5.18}$$

Therefore, substituting β in (5.13) by $\hat{\beta}$ in (5.18) results in the well-known expansion estimator

$$\hat{Y} = \sum_{s} y_k + \hat{\beta} \sum_{c} x_k = N\overline{y}_s \qquad (5.19)$$

This estimator is identical to the common mean model form of the JCE which can be written as

$$\hat{Y}_{JCE} = N\left(\sum_{s_A} y_k + \sum_{s_B} y_k\right) \Big/ \left(n_A + n_B\right) \qquad (5.20)$$

When the probability of sampling duplicates, $k \in s_A \cap s_B$, is ignorable.

Example 5.4.3: ***Linear Regression Model and Linear Regression Estimator***

Under the linear regression model, where the expectation and variance of y_k are

$$\begin{cases} E_{\xi}\left(y_k\right) = \alpha + \beta x_k \\ V_{\xi}\left(y_k\right) = \sigma^2 \end{cases} \qquad (5.21)$$

the BLU of $\hat{Y}$ can be written as the well-known linear egression estimator

$$\hat{Y} = N\left[\overline{y}_s + b\left(\overline{X} - \overline{x}_s\right)\right] \qquad (5.22)$$

where $b = \sum_{s}\left(y_k - \overline{y}_s\right)\left(x_k - \overline{x}_s\right) \Big/ \sum_{s}\left(x_k - \overline{x}_s\right)^2$.

This estimator is identical to the linear regression model form of the JCE which can be written as

$$\hat{Y}_{JCE} = N\left[\frac{\sum_{s_A} y_k + \sum_{s_B} y_k}{n_A + n_B} + b\left(\overline{X} - \frac{\sum_{s_A} x_k + \sum_{s_B} x_k}{n_A + n_B}\right)\right] \qquad (5.23)$$

where $b=\sum_{s}\left(y_k-\frac{\sum_{s_A}y_k+\sum_{s_B}y_k}{n_A+n_B}\right)\left(x_k-\frac{\sum_{s_A}x_k+\sum_{s_B}x_k}{n_A+n_B}\right)\Big/\sum_{s}\left(x_k-\frac{\sum_{s_A}x_k+\sum_{s_B}x_k}{n_A+n_B}\right)^2$ and when the probability of sampling duplicates, $k\in s_A\cap s_B$, is ignorable.

Example 5.4.4: ***Group Mean Model and Stratified Expansion Estimator***

Under the Group Mean Model, where the expectation and variance of y_k are

$$\begin{cases} E_{\xi}\left(y_{jk}\right)=\mu_j \\ V_{\xi}\left(y_{jk}\right)=\sigma_j^2 \end{cases} \tag{5.24}$$

the BLU of $\hat{Y}$ can be written as the well-known stratified expansion estimator

$$\hat{Y}=\sum_{J}N_j\bar{y}_{s_j} \tag{5.25}$$

where s_j denotes the sample cell $U_j\cap s$ with sample size n_j and $\bar{y}_{s_j}=\sum_{s_j}y_j\big/n_j$. Under the ignorable probability of sampling duplicates, $k\in s_A\cap s_B$, this estimator in (5.25) is identical to the group mean model form of the JCE which can be written as

$$\hat{Y}_{JCE}=\sum_{J}\frac{N_j}{\left(n_{Aj}+n_{Bj}\right)}\left(\sum_{s_{Aj}}y_k+\sum_{s_{Bj}}y_k\right) \tag{5.26}$$

where s_{Aj} and s_{Bj} denote the sample cell $U_{Aj}\cap s_A$ and $U_{Bj}\cap s_B$ with sample sizes n_{Aj} and n_{Bj}, respectively.

5.5 Conclusion

As discussed in Chapter 1, the overlap between the dual frame design samples results in a multiplicity problem, in which the overlap domain *ab* includes cases that could be selected twice. The dual frame estimation tries to adjust for this multiplicity problem in the context of the probability sampling theory. Since the prediction theory depends on the relationship between the variables, the dual frame estimation problem has different properties under this theory. In this chapter, the dual frame estimation problem was explored in the context of the prediction theory, in which we found that the dual frame estimation problem is to identify the duplicates and to predict the non-sampled cases.

Similar to the situation for single frame estimation, where the model-based estimators can reduce to well-known design-based estimators, we found that the dual frame model-based estimators can reduce to forms of the JCE. In fact we found that the dual frame estimation problem reduces to a single frame estimation problem and so the general prediction theorem applies. This means that the JCE can be reintroduced as a model-based dual frame estimator, where the general prediction theorem can help in exploring the properties of the JCE under the prediction theory.

Chapter 6

Conclusions and Discussion

6.1 Summary of Dissertation

In this dissertation, the Joint Calibration Estimator (JCE) was proposed as a dual frame estimator that closely meets the desirable properties for the dual frame estimators. The properties of the JCE were discussed through the dissertation chapters. Chapter 1 specified some desirable properties for the dual frame estimators. These properties were discussed for each of the standard dual frame estimators. Chapter 1 concluded with the emergent need for a dual frame estimator that meets these desirable properties. This estimator should be unbiased or approximately so, internally consistent, efficient with low MSE, applicable with standard survey software, robust to non-sampling errors and extendable for multiple frame surveys. Also, it should avoid any unreasonable data or information requirements and should be robust to non-sampling errors in estimator's requirements.

In Chapter 2, the JCE was introduced as an approximately unbiased dual frame estimator. A general expression for the bias resulting from JCE was theoretically derived. This derivation enhanced our understanding and interpretations of the JCE performance. Now it is clear that the performance of the JCE is controlled by the relationship between the study variables and the auxiliary variables, where the best performance happens when the auxiliary variables can interpret the variability in the study variables, or at least when the auxiliary variables are strong correlates of the study variables. Due to the reduced variance of the calibrated estimates, the JCE proved to be an efficient estimator with low MSE. Generally, the efficiency of the JCE depends on how well the variability in the study variables is explained by the auxiliary variables. Since the joint calibration approach results in only one weighting variable, the JCE is an internally consistent estimator that can be applied by standard survey software. Finally, it is straightforward to apply the JCE for the multiple frame surveys estimation.

In Chapter 3, the properties of JCE for dual frame surveys were explored in the presence of the nonresponse error. A general expression for the bias of JCE was theoretically derived; this bias is due to the nature of the nonresponse and the joint calibration approach itself, the latter was derived in Chapter 2. Empirically, the JCE proved to be robust to the nonresponse error as long as a strong set of auxiliary variables is used. This strong set should explain both the response mechanism and the main study variables. Generally, the efficiency of the JCE depends on how well the response mechanism and the variability in the study variables are explained by the auxiliary variables. In the presence of the nonresponse, the JCE can work as both a dual frame estimator that combines the dual frame samples and an adjustment method that adjusts for the nonresponse error.

In Chapter 4, the most unique feature of the JCE was discussed. As opposed to the standard dual frame estimators, the JCE does not require domain membership information. Even if included in the calibration auxiliary variables, the effect of the randomly misclassified domains does not exceed the measurement error effect. Therefore, JCE tends to be robust for the misclassified domains if included in the auxiliary variables. In this chapter, we derived the analytic bias resulting in the standard dual frame estimators from domain misclassification. The misclassification bias expression indicated that within each domain, both the expected total of the study variable for the misclassified cases and the correlation between the misclassification probabilities and the study variable are determinants of the misclassification bias. Therefore, calibrating the data using the auxiliary variables which are correlated with the study variable and the misclassification probabilities could be a promising approach for adjusting the misclassification error.

In Chapter 5, the properties of the dual frame estimation problem were explored under the prediction theory. Since the prediction theory depends on the modeled relationship between the variables, the dual frame estimation problem has different properties under the prediction theory relative to its properties under the probability sampling theory. In this chapter, we found that, under the prediction theory, as long as the duplicates are identifiable, the dual frame estimation problem reduces to a single frame estimation problem and the general prediction theorem for single frame surveys applies. We also found that the model-based dual frame estimators may reduce to the JCE estimators derived in Chapter 2 under the probability sampling theory.

6.2 Future Research and Extensions

There are several extensions to the JCE estimator proposed in this dissertation. For example, in Chapter 2, we derived the JCE for the multiple frame surveys, where the design is composed of more than two frames, but we did not compare the performance of the JCE estimator and the performance of the other dual frame estimators under the multiple frame surveys. For the optimal dual frame estimators, it is difficult to extend the estimator to the multiple frame case. The applicable estimators are extendable but require more domain membership information. In general, under dual frame designs, the JCE achieved comparable results to the standard dual frame estimators with fewer requirements. Under multiple frame designs, the JCE is expected to be a more efficient and practical estimator than the standard estimators.

Another extension is related with the multiplicity information problems discussed in Chapter 4. In this chapter, we only discussed the domain misclassification as an example of the multiplicity information problems. Other problems such as item nonresponse and unknown multiplicity information need to be explored more extensively. Comparisons between the JCE and the standard dual frame estimator in the presence of these problems need to be conducted. Moreover, real information about the magnitude of problems and practical solutions is needed.

Although it is an important objective in most surveys, multiple frame estimation for domains or subpopulations has never been examined in the literature. The properties of the standard dual frame estimators need to be explored for domain estimation. The use of auxiliary information in single frame domain estimation is well documented in the literature (Estevao & Särndal, 1999, 2004; Hidiroglou & Patak, 2004). Therefore, since it depends on accommodating the auxiliary information for dual frame estimation, the joint calibration approach for dual frame domain estimation needs to be examined.

In this dissertation, the JCE was introduced for dual frame estimation. However, in the future the JCE could be extended to be a general approach for combining data from multiple sources. For example, multiple datasets from different surveys could be combined to provide more accurate estimates for study variables that are commonly collected in these surveys. In this case, as long as the calibration auxiliary variables are collected in all surveys, JCE can be easily applied. Finally, extending the ideas in Chapter 5 is necessary to study the properties of the JCE under the prediction theory.

References

Bankier, M.D. (1986). Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Blumberg S.J., & Luke J.V. (2011). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July–December 2010. National Center for Health Statistics. June 2011. Available from: http://www.cdc.gov/nchs/nhis.htm.

Brick, J. M., Brick P.D., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2007). Cell Phone Survey Feasibility in the U.S.: Sampling and Calling Cell Numbers versus Landline Numbers. *Public Opinion Quarterly*, 71:23–39.

Brick, J. M., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2006). Nonresponse Bias in a Dual-frame Sample of Cell and Landline Numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Flores-Cervantes, I., Lee, S., & Norman, G. (2011). Nonsampling Errors in Dual-frame Telephone Surveys. *Survey Methodology*, Vol. 37, No. 1, pp. 1-12.

Casady, R.J., & Lepkowski, J.M. (1993). Stratified Telephone Survey Designs. *Survey Methodology*, 19 (1), 103–13.

Chang, T. & P. S. Kott (2008). Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model. *Biometrika*, 95, 557-571.

Clark, J., Winglee, M. & Liu, B. (2007). Handling Imperfect Overlap Determination in a Dual-frame Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 3233-3238.

Deville, J.C., & Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.C., Särndal, C.E. & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Ekholm, A., & Laaksonen, S. (1991). Weighting via Response Modeling in the Finish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.

Estevao, V.M., & Särndal, C.E. (1999). The Use of Auxiliary Information in Design-based Estimation for Domains. *Survey Methodology*, 25, 213-221.

Estevao, V.M., & Särndal, C.E. (2000). A Functional Form Approach to Calibration. *Journal of Official Statistics*, 16, 379-399.

Estevao, V.M., & Särndal, C.E. (2004). Borrowing Strength is not the Best Technique within a Wide Class of Design-consistent Domain Estimators. *Journal of Official Statistics*, 20, 645-669.

Fuller, W.A., & Burmeister, L.F. (1972). Estimators for Samples Selected from Two Overlapping Frames. *Proceedings of the Social Statistics Section of the American Statistical Association*, 245-249.

González-Villalobos, A., & Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*, Rome: Food and Agriculture Organization of the United Nations. Vols. 1 and 2.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Volume 1.

Hartley, H. O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section of the American Statistical Association*, 203–206.

Hartley, H. O. (1974). Multiple Frame Methodology and Selected Applications. *Sankhya,* Series C, 36, 99–118.

Hidiroglou, M.A., & Patak, P. (2004). Domain Estimation Using Linear Regression. *Survey Methodology*, 30, 67-78.

Horvitz, D.G., & Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, 663-685.

Kalton, G., & Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society,* Ser. A 149, 65-82.

Keeter, S. (2006). The Impact of Cell Phone Noncoverage Bias on Polling in the 2004 Presidential Election. *Public Opinion Quarterly* 70:88–98.

Keeter, S., Kennedy, C., Clark, A., Tompson, T., & Mokrzycki, M. (2007). What's Missing From National Landline RDD Surveys?: The Impact of The Growing Cell-Only Population. *Public Opinion Quarterly* 71:772–792.

Kennedy, C. (2007). Evaluating the Effects of Screening for Telephone Service in Dual-frame RDD Surveys. *Public Opinion Quarterly* 70:750–771.

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

Kott, P.S. & Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse. *Journal of the American Statistical Association*, 105:491, 1265-1275.

Lepkowski, J.M. (1991). Sampling the Difficult to Sample. *Journal of Nutrition*, 121, 416-423.

Lepkowski, J.M., & Groves, R.M. (1986). A Mean Squared Error Model for Multiple Frame, Mixed Mode Survey Design. *Journal of the American Statistical Association*, 81, 930-937.

Lesser, V.M., & Kalsbeek, W.D. (1999). Nonsampling Errors in Environmental Surveys. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*, 4, 473-488.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L., & Mokdad, A.H. (2006). Address-Based Versus Random-Digit Dialed Surveys: Comparison of Key Health and Risk Indicators. *American Journal of Epidemiology,* 164:1019–25.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L., & Mokdad, A.H. (2007). Reaching The U.S Cell Phone Generation: Comparison of Cell Phone Survey Results With an Ongoing Landline Telephone Survey. *Public Opinion Quarterly* 71:814–839.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L., & Mokdad, A.H. (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, *72*, 6-27.

Link, M.W., & Lai, J. (2011). Cell Phone-Only Households and Problems of Differential Nonresponse Using and Address Based Sampling Design. *Public Opinion Quarterly,75(4),* 613-635.

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

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press Publishing Company, Pacific Grove, California.

Lohr, S. (2009). Multiple Frame Surveys. Chapter 4 in D. Pfeffermann & C. R. Rao. (Eds.) Sample Surveys: Design, Methods and Applications, Vol. 29A, (pp. 71-88) Elsevier, The Netherlands: North-Holland.

Lohr, S. (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames. *Survey Methodology*, 37, 197-213.

Lohr, S.L., & Rao, J.N.K. (2000). Inference in Dual Frame Surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., & Rao, J.N.K. (2006). Estimation in Multiple-Frame Surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, Y., & Lohr, S.L. (2010). Gross Flow Estimation in Dual Frame Surveys. *Survey Methodology*, 36, 13-22.

Lündstrom, S., & Särndal, C.E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, 305-327.

Mecatti, F. (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys. *Survey Methodology*, 33, 151-157.

Rao, J.N.K., & Skinner, C.J. (1996), Estimation in Dual Frame Surveys with Complex Designs. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 63-68.

Rao, J.N.K., & Wu, C. (2010). Pseudo-Empirical Likelihood Inference for Dual Frame Surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

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

Särndal, C.E. (2007). The Calibration Approach in Survey Theory and Practice. *Survey Methodology* 33(2):99-119.

Särndal, C.E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.E., & Lundström, S. (2010). Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias. *Survey Methodology* 36(2):131-144.

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

Singh, A.C., & Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. *Survey Methodology*, 22, 107-115.

Singh, A., & Mecatti, F. (2011). Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, 27(4):633–650.

Skinner, C.J. (1991). On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., & Rao, J.N.K. (1996). Estimation in Dual-Frame Surveys with Complex Designs. *Journal of the American Statistical Association*, 91, 349-356.

Stukel, D.M., Hidiroglou, M.A., & Särndal, C.E. (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization. *Survey Methodology*, 22, 117-125.

Sudman, S., & Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.

Valliant, R., Dorfman, A.H., & Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.