

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2026; 11(4): 78-93
© 2026 Stats & Maths
<https://www.mathsjournal.com>
Received: 04-01-2026
Accepted: 06-02-2026

Mutanu Beth
Department of Mathematics and
Actuarial Science, South Eastern
Kenya University, Kitui, Kenya

Kahiri James
Department of Mathematics and
Actuarial Science, Kenyatta
University, Nairobi, Kenya

Odongo Leo
Department of Mathematics and
Actuarial Science, Kenyatta
University, Nairobi, Kenya

Corresponding Author:
Mutanu Beth
Department of Mathematics and
Actuarial Science, South Eastern
Kenya University, Kitui, Kenya

Regression estimation in multiple frames when there are measurement and response errors

Mutanu Beth, Kahiri James and Odongo Leo

DOI: <https://www.doi.org/10.22271/math.2026.v11.i4b.2302>

Abstract

High-quality survey data are fundamental to effective planning and sound decision-making. However, the presence of measurement and response errors in survey data can compromise the validity of statistical analyses by introducing bias and inflating standard errors. Consequently, it is essential to assess and account for the effects of these errors during data analysis to ensure that the resulting information is reliable and suitable for use by statisticians and policymakers. This study focuses on regression estimation within a multiple-frame survey framework. Specifically, a final two-phase sampling design for regression estimation under multiple frames was implemented with the objective of enhancing the precision and robustness of estimates in the presence of measurement and response errors. The findings demonstrate that both types of errors contribute substantially to bias in the estimated parameters. Furthermore, an examination of different error constant values reveals a systematic increase in the magnitude of the estimates as error levels rise, underscoring the pervasive influence of measurement errors on survey outcomes. These results highlight the importance of explicitly incorporating measurement and response error considerations into both the design and implementation of surveys. Failure to do so may lead to misleading inferences and suboptimal policy decisions. Therefore, it is recommended to adopt methodological strategies that account for these errors in order to improve the accuracy, efficiency, and overall reliability of survey-based estimates.

Keywords: Rate of metabolism, blood mass stream rate, warm conductivity, warm era, limited component method, Pennes Bio - Heat Model

Introduction

Multiple frames surveys

Estimation in elusive population in which a direct sampling frame may not exist can be done by use of multiple frames (MF) surveys. Multiple frames are useful for reducing costs for a given precision constraint and improving coverage i.e. under-or over- coverage. To overcome the problem of under coverage the original frame is supplemented with additional frames so that the union of all frames covers the target population. In this case some units may be present in more than one frame, leading to over-coverage. Duplication in the same frame may also occur leading to over-coverage.

Thus imperfect frames can be dealt with by use of MF surveys where two or more listings of units is simultaneously used for sample selection. For instance, we can have a three-frame sampling design by carrying out a MF survey of homeless population where eligibility is from homeless shelters, soup kitchens and street locations.

Thus MF set up corresponds to sampling individually at locations and has been applied in European migration studies. MF surveys allows use of more flexible strategy than conventional single frame survey as well as use of different modes of data collection. For example, in one frame, one can use face-to-face interviews and in another frame, we use email-questionnaires with an aim of controlling coverage survey costs, response rates and estimation accuracy.

MF surveys have significant challenges in efficient estimation of population parameters of interest due to possible overlap among the frames at the selection stage leading to multiple opportunities of selecting the same unit in the final sample.

Therefore, the multiplicity of selected units and their possible duplication at the estimation stage should be dealt with. There is limited use of MF designs though very appealing due to difficulty in a simple interpretation and implementation of MF estimates including variance estimation brought about by the natural complexity of the multiple frame estimation framework. The estimation is difficult due to lack of a standard software. Singh and Mecatti (2009, 2011) ^[15, 14] provided a simplified and unified review of available MF estimation methods by using the multiplicity approach to MF estimation when there no measurement and response errors.

1.1 Estimation in dual frames and multiple frames

The concept of multiple frames arises in surveys when no complete sampling frame comprising all the population units exists. It is assumed that by considering various aspects of the population it is possible to have at least two or more incomplete lists of the population units whose union is the whole population.

As an illustration consider a survey where the sampling frame is those with mail boxes. Not all families may have mail boxes but there are those with mobile phones and no mail boxes and also in the groups there are those with mail boxes and mobile phones.

The list of mail box owners is one frame and the list of mobile telephone owners is another frame.

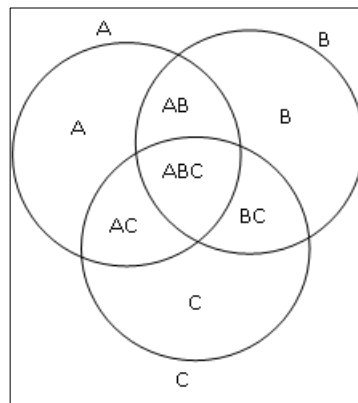
Considering each frame, it is possible to partition the list of units into disjoint sets determined by the membership of the units in other frames. These disjoint sets are called domains of the frame. For instance, if there are two frames say A and B the disjoint sets or the domains of A are two, those in A and those in both A and B i.e. $a(A)$ and $a(AB)$. Similarly, B has two domains $b(B)$ and $b(AB)$. Of course $a(AB)$ and $b(AB)$ are same set.

If there are three frames say A, B, C then A will have four domains as $a(A)$, $a(AB)$, $a(AC)$ and $a(ABC)$. The same with B as $b(B)$, $b(AB)$, $b(BC)$ and $b(ABC)$ and the same with C as $c(C)$, $c(AC)$, $c(BC)$ and $c(ABC)$. In general, if there are say L frames, then frame A will have

$$\binom{L-1}{0} + \binom{L-1}{1} + \dots + \binom{L-1}{L-1} = \sum_{r=0}^{L-1} \binom{L-1}{r} = (1+1)^{L-1} = 2^{L-1}$$

domains. This is the same for all other frames.

The estimation in Multiple Frames (MF) is done by partitioning of all frames into disjoint subsets (Singh and Mecatti, 2014) ^[6]. For example, if we have three frames A, B and C, where A represents the list of people with telephones, B represents the list of people with electricity and C represents the list of people with piped water, the three frames can be partitioned as follows,



The number of domains in a frame is equal to 2^{L-1} and the total number of domains in all the frames is equal to $2^L - 1$

$D_l = 2^{l-1}$ is constant for each frame $l = 1, \dots, L$ in the special case of all domains being non-empty. Define the frame membership indicator $I_{k \in l}, l = 1, \dots, L$ as

$$I_{lk} = \begin{cases} 1 & \text{if } k^{\text{th}} \text{ population unit belongs to frame } l \\ 0, & \text{otherwise} \end{cases}$$

Let m_d be the number of frames a unit in domain d belongs to, where $d = 1, 2, \dots, 2^{L-1}$. Then

$$m_d = \sum_{k \in d} I_{lk}$$

The frame size is denoted by $N_l = \sum_{k=1}^L I_{k \in l}$ note that $\sum_{l=1}^L N_l \geq N$

Consider frame l with 2^{L-1} domains. Each element belonging to domain d appears in m_d frames altogether, which is the multiplicity of such an element. The population mean can therefore be written as,

$$\bar{\mu} = \frac{1}{N} \sum_{l=1}^L \frac{1}{m_d} \sum_{k \in d} \mu_{lk} \quad (1)$$

Now consider that samples are taken one from each frame and denote the samples as S_l from the l^{th} frame. The sample membership can be partitioned into domain membership with frame l as S_{ld} . The estimator is therefore,

$$\hat{\mu} = \frac{1}{N} \sum_{l=1}^L \frac{1}{m_d} \sum_{k \in S_{ld}} \frac{\mu_{lk}}{\pi_{lk}}$$

Where S_d is the sample from domain d and π_{lk} is the inclusion probability of unit k in the sample.

Regression estimation under multiple frames

In regression estimation we assume that different domains have different characteristics and hence the regression estimation will be based on the domains of a frame.

Let the auxiliary variables be p and the observed vector be $(x_{1ldq}, x_{2ldq}, \dots, x_{pldq}) = \underline{x}'_{ldq}$ for the q^{th} unit in domain d of frame L . The true value μ_{ldq} is given by

$$\mu_{ldq} = \underline{x}'_{ldq} \underline{B}_{ldq}, \text{ where } \underline{B}'_{ldq} = (B_{1ldq}, B_{2ldq}, \dots, B_{pldq})$$

The regression estimator is now, $\hat{t}_{\mu_l} = \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{S_{ld}} g_{ldq} \frac{y_{ldq}}{\pi_{ldq}}$, where y_{ldq} is the value observed with,

$$E_m(y_{ldq}) = z_{ldq}, \quad V_m(y_{ldq}) = \sigma_{ldq}^2, \quad Cov_m(y_{ldq}, y_{ldq'}) = \sigma_{ldqq'}^2 \text{ and } g_{ldq} = 1 + \left(\underline{t}_{x_{ld}} - \hat{\underline{t}}_{x_{ld}} \right) \hat{T}^{-1} \frac{x_{ldq}}{\psi_{ldq}^2}$$

The estimator can be written as,

$$\begin{aligned} \hat{t}_{\mu_l} &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} \left[1 + \left(\underline{t}_{x_{ld}} - \hat{\underline{t}}_{x_{ld}} \right) \hat{T}^{-1} \frac{x_{ldq}}{\sigma_{ldq}^2} \right] \frac{y_{ldq}}{\pi_{ldq}} \\ &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \left[\sum_q \frac{y_{ldq}}{\pi_{ldq}} + \left(\underline{t}_{x_{ld}} - \hat{\underline{t}}_{x_{ld}} \right) \hat{T}^{-1} \sum_q \frac{x_{ldq} y_{ldq}}{\sigma_{ldq}^2 \pi_{ldq}} \right] \end{aligned}$$

Taking expectation with respect to the measurement error process we get,

$$\begin{aligned} E_m(\hat{t}_{\mu_l}) &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \left[\sum_{q \in S_{ld}} \frac{z_{ldq}}{\pi_{ldq}} + \left(\underline{t}_{x_{ld}} - \hat{\underline{t}}_{x_{ld}} \right) \hat{T}^{-1} \sum_{S_{ld}} \frac{x_{ldq} z_{ldq}}{\psi_{ldq}^2 \pi_{ldq}} \right] \\ &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} \left[1 + \left(\underline{t}_{x_{ld}} - \hat{\underline{t}}_{x_{ld}} \right) \hat{T}^{-1} \frac{x_{ldq}}{\psi_{ldq}^2} \right] \frac{z_{ldq}}{\pi_{ldq}} = \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} \frac{z_{ldq}}{\pi_{ldq}} \end{aligned} \quad (2)$$

Define the components as

$$\hat{B}_{ldq} = \left(\sum_{q \in S_{ld}} \frac{x_{ldq} x_{ldq'}}{\psi_{ldq}^2} \right)^{-1} \left(\sum_{q \in S_{ld}} \frac{x_{ldq} z_{ldq}}{\psi_{ldq}^2 \pi_{ldq}} \right) \text{ and } \underline{B}_{ldq} = \left(\sum_{q=1}^{N_{ld}} \frac{x_{ldq} x'_{ldq}}{\psi_{ldq}^2} \right)^{-1} \left(\sum_{q=1}^{N_{ld}} \frac{x_{ldq} z_{ldq}}{\psi_{ldq}^2 \pi_{ldq}} \right)$$

Let $z_{ldq}^0 = x_{ldq} B_{ldq}$ and $E_{ldqz} = z_{ldq} - z_{ldq}^0$ then, $E_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d} \left[\sum_{q=1}^{N_{ld}} z_{ldq}^0 + \sum_{q \in S_{ld}} g_{ldq} \frac{E_{ldqz}}{\pi_{ldq}} \right]$

This is approximated under Taylor linearization as,

$$E_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq}^0 + \sum_{q \in S_{ld}} \frac{E_{ldqz}}{\pi_{ldq}}$$

Taking expectation under sampling gives,

$$E_m(\hat{t}_{\mu_i}) = E_p E_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq} \quad (3)$$

The bias introduced by measurement errors is,

$$Bias_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq} - \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} \mu_{ldq} \quad (4)$$

The variance is, $V(\hat{t}_{\mu_i}) = V_p E_m(\hat{t}_{\mu_i}) + E_p V_m(\hat{t}_{\mu_i})$. Under Taylor linearization approximation, the approximate variance is,

$$AV_p E_m(\hat{t}_{\mu_i}) = V_p \sum_d \frac{1}{m_d} \left[\sum_{q=1}^{N_{ld}} z_{ldq}^0 + \sum_{q \in S_{ld}} \frac{E_{ldq}}{\pi_{ldq}} \right]$$

$$AV_p E_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d^2} \left[\sum_q \frac{(1-\pi_{ldq})}{\pi_{ldq}} E_{ldq}^2 + \sum_{q \neq q'} \frac{(\pi_{ldqq'} - \pi_{ldq} \pi_{ldq'})}{\pi_{ldq} \pi_{ldq'}} E_{ldq} E_{ldq'} \right] \quad (5)$$

The response variance component is,

$$E_p V_m(\hat{t}_{\mu_i}) = E_p V_m \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} \frac{y_{ldq}}{\pi_{ldq}} \right]$$

$$= E_p \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} g_{ldq}^2 \frac{\sigma_{ldq}^2}{\pi_{ldq}^2} + \sum_{q \neq q' \in S_{ld}} g_{ldq} g_{ldq'} \frac{\sigma_{ldqq'}^2}{\pi_{ldq} \pi_{ldq'}} \right]$$

With large sample approximation, put $g_{ldq} = 1$, this gives

$$E_p V_m(\hat{t}_{\mu_i}) = E_p \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} \frac{\sigma_{ldq}^2}{\pi_{ldq}^2} + \sum_{q \neq q' \in S_{ld}} \frac{\sigma_{ldqq'}^2}{\pi_{ldq} \pi_{ldq'}} \right] \quad (6)$$

The estimate of the variance is,

$$\hat{V}(\hat{t}_{\mu_i}) = \hat{V} \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} \frac{y_{ldq}}{\pi_{ldq}} \right] = \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} g_{ldq}^2 \frac{y_{ldq}^2}{\pi_{ldq}^2} + \sum_{q \neq q' \in S_{ld}} \frac{g_{ldq} g_{ldq'}}{\pi_{ldq} \pi_{ldq'}} \right] \quad (7)$$

2.1 Special case: Regression estimation under multiple frames

Under simple random sampling without replacement

$$\hat{t}_{\mu_i} = \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{S_{ld}} g_{ldq} y_{ldq} \frac{N_{ld}}{n_{ld}}$$

Which can be written as,

$$\begin{aligned} \hat{t}_{\mu_i} &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} \left[1 + \left(t_{x_{ld}} - \hat{t}_{x_{ld}} \right) \hat{T}^{-1} \frac{x_{ldq}}{\sigma_{ldq}^2} \right] \frac{N_{ld}}{n_{ld}} y_{ldq} = \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \left[\sum_q y_{ldq} \frac{N_{ld}}{n_{ld}} + \left(t_{x_{ld}} - \hat{t}_{x_{ld}} \right) \hat{T}^{-1} \sum_q x_{ldq} y_{ldq} \frac{N_{ld}}{\sigma_{ldq}^2 n_{ld}} \right] \\ \hat{t}_{\mu_i} &= \sum_{d=1}^{2^{L-1}} \frac{N_{ld}}{n_{ld} m_d} \left[\sum_q y_{ldq} + \left(t_{x_{ld}} - \hat{t}_{x_{ld}} \right) \hat{T}^{-1} \sum_q \frac{x_{ldq} y_{ldq}}{\sigma_{ldq}^2} \right] \end{aligned}$$

Taking expectation with respect to the measurement error process we get,

$$\begin{aligned} E_m(\hat{t}_{\mu_i}) &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \left[\sum_{q \in S_{ld}} z_{ldq} \frac{N_{ld}}{n_{ld}} + \left(t_{x_{ld}} - \hat{t}_{x_{ld}} \right) \hat{T}^{-1} \sum_{S_{ld}} x_{ldq} z_{ldq} \frac{N_{ld}}{\psi_{ldq}^2 n_{ld}} \right] \\ &= \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} \left[1 + \left(t_{x_{ld}} - \hat{t}_{x_{ld}} \right) \hat{T}^{-1} \frac{x_{ldq}}{\psi_{ldq}^2} \right] z_{ldq} \frac{N_{ld}}{n_{ld}} = \frac{N_{ld}}{n_{ld}} \sum_{d=1}^{2^{L-1}} \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} z_{ldq} \end{aligned} \quad (8)$$

$$\text{Let } z_{ldq}^0 = x_{ldq} B_{ldq} \text{ and } E_{ldqz} = z_{ldq} - z_{ldq}^0 \text{ then, } E_m(\hat{t}_{lr}) = \sum_d \frac{1}{m_d} \left[\sum_{q=1}^{N_{ld}} z_{ldq}^0 + \sum_{q \in S_{ld}} g_{ldq} E_{ldqz} \frac{N_{ld}}{n_{ld}} \right]$$

This is approximated under Taylor linearization as,

$$E_m(\hat{t}_{lr}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq}^0 + \frac{N_{ld}}{n_{ld}} \sum_{q \in S_{ld}} E_{ldqz}$$

Taking expectation under sampling gives,

$$E_m(\hat{t}_{lr}) = E_p E_m(\hat{t}_{\mu_i}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq} \quad (9)$$

The bias introduced by measurement errors is,

$$Bias_m(\hat{t}_{lr}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} z_{ldq} - \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{ld}} \mu_{ldq} \quad (10)$$

The variance is, $V(\hat{t}_{lr}) = V_p E_m(\hat{t}_{lr}) + E_p V_m(\hat{t}_{lr})$. Under Taylor linearization approximation, the approximate variance is,

$$\begin{aligned} AV_p E_m(\hat{t}_{lr}) &= V_p \sum_d \frac{1}{m_d} \left[\sum_{q=1}^{N_{ld}} z_{ldq}^0 + \sum_{q \in S_{ld}} E_{ldq} \frac{N_{ld}}{n_{ld}} \right] \\ &= \sum_d \frac{1}{m_d^2} \left[\sum_q \frac{(N_{ld} - n_{ld})}{n_{ld}} E_{ldq}^2 + \sum_{q \neq q'} \left(\frac{N'_{ld}(n_{ld} - 1)}{n'_{ld}(N'_{ld} - 1)} - 1 \right) E_{ldq} E_{ldq'} \right] \end{aligned} \quad (11)$$

The response variance component is,

$$E_p V_m(\hat{t}_{lr}) = E_p V_m \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} y_{ldq} \frac{N_{ld}}{n_{ld}} \right]$$

$$= E_p \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} g_{ldq}^2 \sigma_{ldq}^2 \frac{N_{ld}^2}{n_{ld}^2} + \sum_{q \neq q' \in S_{ld}} g_{ldq} g_{ldq'} \sigma_{ldqq'}^2 \frac{N_{ld} N'_{ld}}{n_{ld} n'_{ld}} \right]$$

With large sample approximation, put $g_{ldq} = 1$, this gives

$$E_p V_m(\hat{t}_{lr}) = E_p \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} \sigma_{ldq}^2 \frac{N_{ld}^2}{n_{ld}^2} + \sum_{q \neq q' \in S_{ld}} \sigma_{ldqq'}^2 \frac{N_{ld} N'_{ld}}{n_{ld} n'_{ld}} \right]$$

The estimate of the variance is,

$$\hat{V}(\hat{t}_{lr}) = \hat{V} \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{ld}} g_{ldq} y_{ldq} \frac{N_{ld}}{n_{ld}} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} g_{ldq}^2 y_{ldq}^2 \frac{N_{ld}^2}{n_{ld}^2} \sum_{q \neq q'} \sum_{q' \in S_{ld}} g_{ldq} g_{ldq'} \frac{N_{ld} N'_{ld}}{n_{ld} n'_{ld}} \right]$$

Thus,

$$\hat{V}(\hat{t}_{lr}) = \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{ld}} \frac{N_{ld}^2}{n_{ld}^2} y_{ldq}^2 \sum_{q \neq q'} \sum_{q' \in S_{ld}} \frac{N_{ld} N'_{ld}}{n_{ld} n'_{ld}} y_{ldq} y'_{ldq} \right] \quad (12)$$

Two-phase regression estimation under multiple frames with errors

Let N_{fd} be the size of domain d in frame f . In two-phase, an initial sample S_{1fd} is selected say of size n_{fdq} . For the selected units, observe the vector of auxiliary variable $\mathbf{x}'_{fdq} = (x_{1fdq}, x_{2fdq}, \dots, x_{pfdq})$. The q^{th} unit is selected with probability π_{fdq} . The joint inclusion probability of two units q and q' is $\pi_{1fdqq'}$. In the 2^{nd} phase, select sample S_{fd} from S_{1fd} and observe y_{fdq} . The inclusion probability is π_{fdq/s_1} and the joint inclusion probability is $\pi_{fdqq'/s_{fd}}$. The overall probability of including unit q in final sample S_{fd} is, $\pi_{fdq} = \pi_{1fdq} \pi_{fdq/s_1}$, also $\pi_{fdqq'} = \pi_{1fdqq'} \pi_{fdqq'/s_{fd}}$.

$E_m(y_{fdq}) = Z_{fdq}$, $V_m(y_{fdq}) = \sigma_{fdq}^2$ and $Cov(y_{fdq}, y_{fdq'}) = \sigma_{fdqq'}$. Let $\hat{y}_{fdq} = \mathbf{x}'_{fdq} \hat{\mathbf{B}}_{fd}$. The regression estimator of the frame f total can be written as, $\hat{t}_{\mu_f} = \sum_d \frac{1}{m_d} \left[\sum_{S_{1fd}} \frac{\hat{y}_{fdq}}{\pi_{1fdq}} + \sum_{S_{fd}} \frac{y_{fdq} - \hat{y}_{fdq}}{\pi_{fdq}} \right]$

This estimator can be written as,

$$\hat{t}_{\mu_f} = \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} \frac{y_{fdq}}{\pi_{fdq}} + \sum_{q \in S_{1fd}} \frac{\hat{y}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \frac{\hat{y}_{fdq}}{\pi_{fdq}} \right]$$

$$= \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} \frac{y_{fdq}}{\pi_{fdq}} + \left(\sum_{q \in S_{1fd}} \frac{\mathbf{x}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \frac{\mathbf{x}_{fdq}}{\pi_{fdq}} \right) \left(\sum_{q \in S_{fd}} \frac{\mathbf{x}'_{fdq} \mathbf{x}_{fdq'}}{\psi_{fdq}^2} \right)^{-1} \sum_{q \in S_{fd}} \frac{\mathbf{x}_{fdq} y_{fdq}}{\psi_{fdq}^2 \pi_{fdq}} \right]$$

$$\begin{aligned}
&= \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} \frac{y_{fdq}}{\pi_{fdq}} + \left(\sum_{q \in S_{1fd}} \frac{\tilde{x}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq}}{\pi_{fdq}} \right) \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq} y_{fdq}}{\psi_{fdq}^2 \pi_{fdq}} \right] \\
&= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} \left[1 + \left(\sum_{q \in S_{1fd}} \frac{\tilde{x}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq}}{\pi_{fdq}} \right) \hat{T}^{-1} \frac{\tilde{x}_{fdq}}{\psi_{fdq}^2} \right] \frac{y_{fdq}}{\pi_{fdq}}
\end{aligned}$$

The expectation of the estimator is now based on three levels, namely, first phase, second phase and the measurement error model phase. The expectation of the estimator is therefore,

$$E(\hat{t}_{\mu_f}) = E_1 E_2 E_m(\hat{t}_{\mu_f})$$

Under the measurement error model,

$$\begin{aligned}
E_m(\hat{t}_{\mu_f}) &= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} \left[1 + \left(\sum_{q \in S_{1fd}} \frac{\tilde{x}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq}}{\pi_{fdq}} \right) \hat{T}^{-1} \frac{\tilde{x}_{fdq}}{\psi_{fdq}^2} \right] \frac{Z_{fdq}}{\pi_{fdq}} \\
&= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} g_{fdq} \frac{Z_{fdq}}{\pi_{fdq}}
\end{aligned} \tag{13}$$

Define, $Z_{fdq}^0 = \tilde{x}'_{fdq} \tilde{B}_{fdz}$ where, $\tilde{B}_{fdz} = \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq} Z_{fdq}}{\psi_{fdq}^2 \pi_{1fdq}}$ and $\hat{\tilde{B}}_{fdz} = \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{\tilde{x}_{fdq} Z_{fdq}}{\psi_{fdq}^2 \pi_{fdq}}$

And $\hat{Z}_{fdq} = \tilde{x}'_{fdq} \hat{\tilde{B}}_{fdz}$. Then define $E_{fdqz} = \hat{Z}_{fdq} - Z_{fdq}^0$, the estimator now can be written as

$$E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{Z_{fdq}^0}{\pi_{1fdq}} + \sum_{q \in S_{fd}} g_{fdq} \frac{E_{fdqz}}{\pi_{fdq}} \right]$$

Under Taylor linearization approximation, $E_m(\hat{t}_{\mu_f})$ can be written as,

$$E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{Z_{fdq}^0}{\pi_{1fdq}} + \sum_{q \in S_{fd}} \frac{E_{fdqz}}{\pi_{fdq}} \right]$$

Taking expectation under the 2^{nd} phase now gives,

$$\begin{aligned}
E_2 E_m(\hat{t}_{\mu_f}) &= E_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{Z_{fdq}^0}{\pi_{1fdq}} + \sum_{q \in S_{fd}} \frac{\hat{Z}_{fdq} - Z_{fdq}^0}{\pi_{fdq}} \right] = \sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{Z_{fdq}}{\pi_{1fdq}} \text{ and} \\
E(\hat{t}_{\mu_f}) &= E_1 E_2 E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{fd}} Z_{fdq}
\end{aligned} \tag{14}$$

Again to get the approximated variance,

$$V(\hat{t}_{\mu_f}) = V_1 E_2 E_m(\hat{t}_{\mu_f}) + E_1 V_2 E_m(\hat{t}_{\mu_f}) + E_1 E_2 V_m(\hat{t}_{\mu_f})$$

$$V_1 E_2 E_m(\hat{t}_{\mu_f}) = V_1 E_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{Z_{fdq}^0}{\pi_{1fdq}} + \sum_{q \in S_{fd}} \frac{E_{fdqz}}{\pi_{fdq}} \right] = V_1 \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{Z_{fdqz}}{\pi_{1fdq}} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q=1}^{N_{fd}} \frac{(1-\pi_{fdq})}{\pi_{fdq}} Z_{fdq}^2 + \sum_{q \neq q'} \frac{\pi_{1fdq} - \pi_{1fdq'} \pi_{1fdq'}}{\pi_{1fdq} \pi_{1fdq'}} Z_{fdq} Z_{fdq'} \right] \quad (15)$$

$$\begin{aligned} E_1 V_2 E_m(\hat{t}_{\mu_f}) &= E_1 V_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{Z_{fdq}^0}{\pi_{1fdq}} + \sum_{q \in S_{fd}} \frac{E_{fdq}}{\pi_{fdq}} \right] = E_1 V_2 \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{E_{fdqz}}{\pi_{fdq}} I_{2fdq} \right] \\ &= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{1fd}} \frac{\pi_{fdq/S_1} (1-\pi_{fdq/S_1})}{\pi_{fdq}^2} E_{fdqz}^2 + \sum_{q \neq q' \in S_{1fd}} \frac{\pi_{fdq/S_1} - \pi_{fdq/S_1} \pi_{fdq'/S_1}}{\pi_{fdq} \pi_{fdq'}} E_{fdqz} E_{fdq'z} \right] \\ &= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{1fd}} \frac{\pi_{fdq/S_1} (1-\pi_{fdq/S_1})}{\pi_{fdq}^2} E_{fdqz}^2 I_{1fdq} + \sum_{q \neq q' \in S_{1fd}} \frac{\pi_{fdq/S_1} - \pi_{fdq/S_1} \pi_{fdq'/S_1}}{\pi_{fdq} \pi_{fdq'}} E_{fdqz} E_{fdq'z} I_{fdqq'} \right] \\ &= \sum_d \frac{1}{m_d^2} \left[\sum_{q=1}^{N_{fd}} \frac{(1-\pi_{fdq/S_1})}{\pi_{fdq}} E_{fdq}^2 + \sum_{q \neq q'} \frac{\pi_{fdq} (\pi_{fdq'} - \pi_{fdq/S_1} \pi_{fdq'/S_1})}{\pi_{fdq} \pi_{fdq'}} E_{fdqz} E_{fdq'z} \right] \quad (16) \end{aligned}$$

To get the component $E_1 E_2 V_m(\hat{t}_{\mu_f})$, write

$$\begin{aligned} V_m(\hat{t}_{\mu_f}) &= V_m \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} g_{fdq} \frac{y_{fdq}}{\pi_{fdq}} \right] \\ &= \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} g_{fdq}^2 \frac{\sigma_{fdq}^2}{\pi_{fdq}^2} + \sum_{q \neq q'} g_{fdq} g_{fdq'} \frac{\sigma_{fdqq'}}{\pi_{fdq} \pi_{fdq'}} \right] \end{aligned}$$

Under Taylor linearization approximation, $g_{fdq} = 1$

$$\begin{aligned} \Rightarrow E_1 E_2 V_m(\hat{t}_{\mu_f}) &= E_1 E_2 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{\sigma_{fdq}^2}{\pi_{fdq}^2} + \sum_{q \neq q'} \frac{\sigma_{fdqq'}}{\pi_{fdq} \pi_{fdq'}} \right] \\ &= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{\pi_{fdq/S_1} \sigma_{fdq}^2}{\pi_{fdq}^2} + \sum_{q \neq q'} \frac{\pi_{fdq/S_1} \sigma_{fdqq'}}{\pi_{fdq} \pi_{fdq'}} \right] \\ &= \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{\sigma_{fdq}^2}{\pi_{fdq}} + \sum_{q \neq q'} \frac{\pi_{fdq} \sigma_{fdqq'}}{\pi_{fdq} \pi_{fdq'}} \right] \quad (17) \end{aligned}$$

3.1 Special case: Two-phase regression estimation under multiple frames with errors

The regression estimator of the frame f total under simple random sampling without replacement can be written as,

$$\hat{t}_{\mu_f} = \sum_d \frac{1}{m_d} \left[\sum_{S_{1fd}} \hat{y}_{fdq} \frac{N_{1fd}}{n_{1fdq}} + \sum_{S_{fd}} (y_{fdq} - \hat{y}_{fdq}) \frac{N_{fd}}{n_{fd}} \right]$$

This estimator can be written as,

$$\begin{aligned}
\hat{t}_{\mu_f} &= \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} y_{fdq} \frac{N_{fd}}{n_{fd}} + \sum_{q \in S_{1fd}} \hat{y}_{fdq} \frac{N_{1fd}}{n_{1fd}} - \sum_{q \in S_{fd}} \hat{y}_{fdq} \frac{N_{fd}}{n_{fd}} \right] \\
&= \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} y_{fdq} \frac{N_{fd}}{n_{fd}} + \left(\sum_{q \in S_{1fd}} \tilde{x}_{fdq} \frac{N_{1fd}}{n_{1fd}} - \sum_{q \in S_{fd}} \tilde{x}_{fdq} \frac{N_{fd}}{n_{fd}} \right) \left(\sum_{q \in S_{fd}} \frac{\tilde{x}'_{fdq} \tilde{x}_{fdq}}{\psi_{fdq}^2} \right)^{-1} \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \frac{\tilde{x}_{fdq} y_{fdq}}{\psi_{fdq}^2} \right] \\
&= \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{fd}} y_{fdq} \frac{N_{fd}}{n_{fd}} + \left(\sum_{q \in S_{1fd}} \tilde{x}_{fdq} \frac{N_{1fd}}{n_{1fd}} - \sum_{q \in S_{fd}} \tilde{x}_{fdq} \frac{N_{fd}}{n_{fd}} \right) \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \frac{\tilde{x}_{fdq} y_{fdq}}{\psi_{fdq}^2} \right] \\
&= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} \left[1 + \left(\sum_{q \in S_{1fd}} \tilde{x}_{fdq} \frac{N_{1fd}}{n_{1fd}} - \sum_{q \in S_{fd}} \tilde{x}_{fdq} \frac{N_{fd}}{n_{fd}} \right) \hat{T}^{-1} \frac{\tilde{x}_{fdq}}{\psi_{fdq}^2} \right] \frac{N_{fd}}{n_{fd}} y_{fdq}
\end{aligned}$$

Thus,

$$\hat{t}_{\mu_f} = \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} \left[1 + \left(\sum_{q \in S_{1fd}} \frac{N_{1fd}}{n_{1fd}} \tilde{x}_{fdq} - \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \tilde{x}_{fdq} \right) \hat{T}^{-1} \frac{\tilde{x}_{fdq}}{\psi_{fdq}^2} \right] \frac{N_{fd}}{n_{fd}} y_{fdq}$$

The expectation of the estimator is therefore,

$$E(\hat{t}_{\mu_f}) = E_1 E_2 E_m(\hat{t}_{\mu_f})$$

Under the measurement error model,

$$\begin{aligned}
E_m(\hat{t}_{\mu_f}) &= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} \left[1 + \left(\sum_{q \in S_{1fd}} \tilde{x}_{fdq} \frac{\tilde{x}_{fdq}}{\pi_{1fdq}} - \sum_{q \in S_{fd}} \tilde{x}_{fdq} \frac{\tilde{x}_{fdq}}{\pi_{fdq}} \right) \hat{T}^{-1} \frac{\tilde{x}_{fdq}}{\psi_{fdq}^2} \right] \frac{Z_{fdq}}{\pi_{fdq}} \\
&= \sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} g_{fdq} Z_{fdq} \frac{N_{fd}}{n_{fd}}
\end{aligned} \tag{18}$$

Define, $Z_{fdq}^0 = \tilde{x}'_{fdq} \tilde{B}_{fdz}$ where, $\tilde{B}_{fdz} = \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{N_{1fd}}{n_{1fd}} \frac{\tilde{x}_{fdq} Z_{fdq}}{\psi_{fdq}^2}$ and $\hat{B}_{fdz} = \hat{T}^{-1} \sum_{q \in S_{fd}} \frac{N_{1fd}}{n_{1fd}} \frac{\tilde{x}_{fdq} Z_{fdq}}{\psi_{fdq}^2}$

And $\hat{Z}_{fdq} = \tilde{x}'_{fdq} \hat{B}_{fdz}$. Then define $E_{fdqz} = \hat{Z}_{fdq} - Z_{fdq}^0$, the estimator now can be written as

$$E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} Z_{fdq}^0 \frac{N_{1fd}}{n_{1fd}} + \sum_{q \in S_{fd}} g_{fdq} E_{fdqz} \frac{N_{fd}}{n_{fd}} \right]$$

Under Taylor linearization approximation, $E_m(\hat{t}_{\mu_f})$ can be written as,

$$E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} Z_{fdq}^0 \frac{N_{1fd}}{n_{1fd}} + \sum_{q \in S_{fd}} E_{fdqz} \frac{N_{fd}}{n_{fd}} \right]$$

Taking expectation under the 2^{nd} phase now gives,

$$E_2 E_m(\hat{t}_{\mu_f}) = E_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} Z_{fdq}^0 \frac{N_{1fd}}{n_{1fd}} + \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} (\hat{Z}_{fdq} - Z_{fdq}^0) \right] = \sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{N_{1fd}}{n_{1fd}} Z_{fdq} \text{ and}$$

$$E(\hat{t}_{\mu_f}) = E_1 E_2 E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d} \sum_{q=1}^{N_{fd}} Z_{fdq} \quad (19)$$

To get the approximated variance,

$$V(\hat{t}_{\mu_f}) = V_1 E_2 E_m(\hat{t}_{\mu_f}) + E_1 V_2 E_m(\hat{t}_{\mu_f}) + E_1 E_2 V_m(\hat{t}_{\mu_f})$$

$$V_1 E_2 E_m(\hat{t}_{\mu_f}) = V_1 E_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{N_{1fd}}{n_{1fd}} Z_{fdq}^0 + \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} E_{fdqz} \right] = V_1 \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{N_{1fd}}{n_{1fd}} Z_{fdqz} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q=1}^{N_{fd}} \left(\frac{N_{fd}}{n_{fd}} - 1 \right) Z_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd} (n_{fd} - 1) n'_{fd} (N_{fd} - 1)}{N'_{fd} (N_{fd} - 1)} Z_{fdq} Z_{fdq'} \right] \quad (20)$$

$$E_1 V_2 E_m(\hat{t}_{\mu_f}) = E_1 V_2 \sum_d \frac{1}{m_d} \left[\sum_{q \in S_{1fd}} \frac{N_{1fd}}{n_{1fd}} Z_{fdq}^0 + \sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} E_{fdq} \right] = E_1 V_2 \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{1fd}} \frac{N_{fd}}{n_{fd}} E_{fdqz} I_{2fdq} \right]$$

$$= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{1fd}} \frac{\pi_{fdq/S_1} (N_{fd}^2 - N_{fd} n_{fd})}{n_{fd}^2} E_{fdqz}^2 + \sum_{q \neq q' \in S_{1fd}} \left[\frac{N'_{fd} (n_{fd} - 1)}{n'_{fd} (N_{fd} - 1)} - 1 \right] E_{fdqz} E_{fdq'z} \right]$$

$$= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{1fd}} \left(\frac{N_{fd}}{n_{fd}} - 1 \right) E_{fdqz}^2 I_{1fdq} + \sum_{q \neq q' \in S_{1fd}} \left(\frac{N'_{fd} (n_{fd} - 1)}{n'_{fd} (N_{fd} - 1)} - 1 \right) E_{fdqz} E_{fdq'z} I_{fdq q'} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q=1}^{N_{fd}} \left(\frac{N_{fd}}{n_{fd}} - 1 \right) E_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd} (n_{fd} - 1) n'_{fd} (N_{fd} - 1)}{N'_{fd} (N_{fd} - 1)} E_{fdqz} E_{fdq'z} \right]$$

Thus,

$$E_1 V_2 E_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d^2} \left[\sum_{q=1}^{N_{fd}} \left(\frac{N_{fd}}{n_{fd}} - 1 \right) E_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd} (n_{fd} - 1) n'_{fd} (N_{fd} - 1)}{N'_{fd} (N_{fd} - 1)} E_{fdqz} E_{fdq'z} \right] \quad (21)$$

To get the component $E_1 E_2 V_m(\hat{t}_{\mu_f})$, write

$$V_m(\hat{t}_{\mu_f}) = V_m \left[\sum_d \frac{1}{m_d} \sum_{q \in S_{fd}} g_{fdq} y_{fdq} \frac{N_{fd}}{n_{fd}} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} g_{fdq}^2 \sigma_{fdq}^2 \frac{N_{fd}^2}{n_{fd}^2} + \sum_{q \neq q'} g_{fdq} g_{fdq'} \sigma_{fdq q'} \frac{N_{fd} N'_{fd}}{n_{fd} n'_{fd}} \right]$$

Under Taylor linearization approximation, $g_{fdq} = 1$

$$\Rightarrow E_1 E_2 V_m(\hat{t}_{\mu_f}) = E_1 E_2 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \sigma_{fdq}^2 \frac{N_{fd}^2}{n_{fd}^2} + \sum_{q \neq q'} \frac{N_{fd} N'_{fd}}{n_{fd} n'_{fd}} \sigma_{fdq q'} \right]$$

$$= E_1 \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \sigma_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd}(n_{fd}-1)}{n'_{fd}(N_{fd}-1)} \sigma_{fdqq'} \right]$$

$$= \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \sigma_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd}(n_{fd}-1)}{n'_{fd}(N_{fd}-1)} \sigma_{fdqq'} \right]$$

Therefore,

$$E_1 E_2 V_m(\hat{t}_{\mu_f}) = \sum_d \frac{1}{m_d^2} \left[\sum_{q \in S_{fd}} \frac{N_{fd}}{n_{fd}} \sigma_{fdq}^2 + \sum_{q \neq q'} \frac{N'_{fd}(n_{fd}-1)}{n'_{fd}(N_{fd}-1)} \sigma_{fdqq'} \right] \quad (22)$$

4. Results and Discussion

4.1 Regression estimation under multiple frames

The table below represents the distribution of units of the generated population 300 with the specified regression parameters and population size for each frame.

Table 1: Distribution of units of the generated population 300

Frame	domains in that Frame	Number of Units
Frame_1	a,ab,ac,abc	126
Frame_2	b,ab,bc,abc	100
Frame_3	c, ac,bc,abc	114
Total		340

X represents the auxiliary variable. Y is the variable of interest, modeled as a linear function of X with some random error. In frame 1 the variable X has an average of 60.19399779 while Y has an average of 185.5474753. Frame 1 has a total of 126 units. In frame 2 the variable X has an average of 59.75562628 while Y has an average of 184.3415139. Frame2 has a total of 100 units. In frame 3 the variable X has an average of 59.58270185 while Y has an average of 183.9812518. Frame 3 has a total of 114 units.

4.2 Single-Phase Regression Estimation under Multiple Frames

Errors are introduced to the Y values based on the specified k and p values. Results are stored in data frames for each combination of k and p. Regression estimates and mean/variance for the error-introduced data for each combination of k and p are computed.

The table below provides the totals, biases, and variances of the estimated regression parameters b_{10} and b_{11} under different error conditions.

Table 2: Single-phase estimation of parameters in frame 1 with varying values of k and p

k	P	$\hat{b}_{10}$	$\hat{b}_{11}$	$\hat{t}_{1dq}$	b_{10} bias	b_{11} bias	b_{10} var	b_{11} var
0.01	0.01	21.2848	2.7290	23379.0383	16.2848	-0.2710	42.1402	0.0116
0.01	0.05	21.2323	2.7290	23373.4524	16.2323	-0.2710	43.2084	0.0119
0.05	0.01	20.2215	2.7465	23378.6511	15.2215	-0.2535	43.7447	0.0120
0.05	0.05	21.1927	2.7287	23366.7765	16.1927	-0.2713	42.3904	0.0117
0.1	0.01	22.5074	2.7067	23378.9868	17.5074	-0.2933	42.5119	0.0117
0.1	0.05	21.9053	2.7174	23373.9234	16.9053	-0.2826	44.0561	0.0121
0.2	0.01	31.7229	2.5562	23378.4220	26.7230	-0.4438	44.6037	0.0123
0.2	0.05	29.2231	2.5969	23375.5634	24.2231	-0.4031	44.8328	0.0123

Table 3: Single-phase estimation of parameters in frame 2 with varying values of k and p

k	P	$\hat{b}_{20}$	$\hat{b}_{21}$	$\hat{t}_{2dq}$	b_{10} bias	b_{21} bias	b_{20} var	b_{21} var
0.01	0.01	14.5873	2.8406	18432.6897	9.5873	-0.1594	37.4096	0.0104
0.01	0.05	14.1979	2.8461	18426.2538	9.1980	-0.1539	38.2844	0.0107
0.05	0.01	15.6431	2.8242	18436.4282	10.6431	-0.1758	38.2077	0.0107
0.05	0.05	14.3395	2.8420	18422.9978	9.3395	-0.1580	39.4647	0.0110
0.1	0.01	19.1807	2.7643	18435.7682	14.1807	-0.2357	37.0760	0.0104
0.1	0.05	15.0566	2.8351	18431.1144	10.0566	-0.1649	37.0621	0.0104
0.2	0.01	22.1185	2.7106	18433.1850	17.1185	-0.2894	41.7301	0.0116
0.2	0.05	20.1649	2.7546	18444.4755	15.1649	-0.2454	55.8658	0.0157

Table 4: Single-phase estimation of parameters in frame 3 with varying values of k and p

k	P	$\hat{b}_{30}$	$\hat{b}_{31}$	$\hat{t}_{3dq}$	b_{30} bias	b_{31} bias	b_{30} var	b_{31} var
0.01	0.01	13.9710	2.8535	20974.6391	8.9710	-0.1465	38.4900	0.0108
0.01	0.05	14.5072	2.8439	20971.9109	9.5072	-0.1561	39.7591	0.0112
0.05	0.01	15.3330	2.8312	20972.8491	10.3330	-0.1688	37.5907	0.0106
0.05	0.05	12.1810	2.8825	20967.5725	7.1810	-0.1175	40.2686	0.0113
0.1	0.01	14.8036	2.8389	20974.4919	9.8036	-0.1611	40.9844	0.0115
0.1	0.05	18.7959	2.7690	20963.7899	13.7959	-0.2310	47.9199	0.0134
0.2	0.01	17.5771	2.7936	20973.8178	12.5771	-0.2064	50.1227	0.0141
0.2	0.05	20.8755	2.7404	20976.1125	15.8755	-0.2596	48.9794	0.0138

b_0 : The bias is consistently high (15–26) across all parameter combinations, with the highest values at higher k values e.g., 26.72 for k=0.2, p=0.01. This suggests the model systematically overestimates the intercept, especially for larger k.

b_1 : The bias is negative (-0.27 to -0.44), indicating underestimation of the true slope. The magnitude increases with k e.g., -0.44 for k=0.2, p=0.01, showing worse performance at higher sampling fractions.

b_0 : Ranges between 42–45, with minimal sensitivity to k or p. The variance is stable but high, reflecting significant uncertainty in intercept estimates.

b_1 : Small (0.011–0.012) and stable, indicating precise slope estimates despite bias.

b_0 : Starts high (16) at k=0.01 and increases sharply with k, peaking at 26.7 for k=0.2.

The bias is worse for p=0.01 than p=0.05 at higher k, suggesting sensitivity to both parameters.

b_1 : Stable but negative (-0.27 to -0.44), indicating systematic underestimation of the slope. Bias worsens with k e.g., -0.44 at k=0.2, especially for p=0.01.

b_0 Variance: High (42–45) and insensitive to k, but slightly higher for k=0.2.

This implies uncertainty in intercept estimates is consistently large, regardless of sampling.

b_1 Variance: Very low (0.011–0.012) and flat, indicating precise but biased slope estimates.

b_0 : Bias ranges from 9.2 to 17.1, increasing with k e.g., 17.12 for k=0.2, p=0.01. Overestimation is systematic but less severe than Frame 1.

b_1 : Bias is negative (-0.16 to -0.29), with larger biases at higher k. The slope is consistently underestimated but less so than in Frame 1.

b_0 Variance: Lower (37–56) than Frame 1, except for the combination k=0.2, p=0.05 (55.87), which shows instability.

b_1 Variance: Slightly higher than Frame 1 (0.010–0.016), with a notable jump for k=0.2, p=0.05 (0.0157), indicating reduced precision at high k and p.

b_0 : Lower than Frame 1 (9–17) but still increases with k. Notable spike at k=0.2 e.g., 17.1 for p=0.01.

b_1 : Negative (-0.16 to -0.29) but less severe than Frame 1. Bias is stable for $k \leq 0.1$ but worsens at k=0.2.

b_0 : Lower than Frame 1 (~37–56), but spikes at k=0.2, p=0.05. This suggests instability with aggressive sampling ('k=0.2').

b_1 : Slightly higher than Frame 1 (0.010–0.016), with a sharp rise at k=0.2, p=0.05.

b_0 , b_1 and t_{ldq} estimates for Frame 1, 2, and 3

b_0 : Bias ranges from 7.2 to 15.9, increasing with k, e.g., 15.88 for k=0.2, p=0.05. Overestimation is present but less extreme than Frames 1–2.

b_1 : Bias is negative (-0.12 to -0.26), with minimal sensitivity to p. The slope is slightly underestimated, but the bias is the smallest among all frames.

b_0 : Ranges from 38–50, increasing with k, e.g., 50.12 for k=0.2, p=0.01. Higher k leads to more unstable intercept estimates.

b_1 : Small (0.010–0.014) and stable, similar to other frames but with the lowest bias.

b_0 : Smallest among frames (~7–16) but still increases with k. Best performance at k=0.05, p=0.05 (bias = 7.2).

b_1 : Mildly negative (-0.12 to -0.26), with the least bias across frames. Stable for $k \leq 0.1$ but deteriorates slightly at 'k=0.2'.

b_0 : Moderate (38–50) and scales with k, but no extreme spikes.

b_1 : Lowest of all frames (0.010–0.014), with minimal sensitivity to k or p.

$\hat{b}_0$: Increases significantly with k, e.g., from 21 at k=0.01 to 31.7 at k=0.2).

The intercept is overestimated, especially at higher sampling fractions k , indicating a systematic upward bias. The worst overestimation occurs at $k=0.2$.

$\hat{b}_1$: Decreases with k , e.g., from 2.73 at $k=0.01$ to 2.56 at $k=0.2$.

The slope is underestimated, with the bias worsening as k increases

$\hat{t}_{dq}$ Stable across k and p , with minimal fluctuations.

The mean estimate is robust to changes in k and p , suggesting the sampling strategy does not heavily distort the central tendency.

$\hat{b}_0$ starts high (21) at $k=0.01$ and increases dramatically to 32 at $k=0.2$. Shows clear positive relationship with k , indicating severe overestimation that worsens with sampling fraction. $p=0.01$ line rises more steeply than $p=0.05$ at higher k values.

$\hat{b}_1$ begins near 2.73 at $k=0.01$ and declines to 2.56 at $k=0.2$. Demonstrates consistent negative trend with increasing k . Both p values follow similar downward trajectories

$\hat{t}_{dq}$ remains extremely stable around 23,370-23,380 across all k values. Shows virtually no sensitivity to either k or p changes.

$\hat{b}_0$ increases with k but less severely than Frame 1, e.g., 14.2 at $k=0.01$ to 22.1 at $k=0.2$.

Intercept bias exists but is less extreme than Frame 1. Higher k still leads to overestimation.

$\hat{b}_1$ mildly decreases with k , e.g., 2.84 at $k=0.01$ to 2.71 at $k=0.2$. Slope estimates are closer to the truth than Frame 1 but still exhibit slight underestimation at higher k .

$\hat{t}_{dq}$ very stable, with negligible sensitivity to k or p . Like Frame 1, the mean estimate is reliable across sampling strategies.

$\hat{b}_0$ ranges from 14 ($k=0.01$) to 22 ($k=0.2$). Shows moderate positive relationship with k . $p=0.01$ produces higher intercepts than $p=0.05$ at same k values

$\hat{b}_1$ declines gently from 2.84 to 2.71 across k values. Maintains better stability than Frame 1. $p=0.05$ shows slightly less bias than $p=0.01$

$\hat{t}_{dq}$ hovers consistently around 18,430 and shows remarkable stability. Tiny variations (<0.1) across conditions.

$\hat{b}_0$ increases with k , e.g., 14 at $k=0.01$ to 20.9 at $k=0.2$ but is the least biased of all frames. Intercept estimates are closest to the true value, though still slightly overestimated at higher k .

$\hat{b}_1$ nearly stable, with only minor decreases at higher k , e.g., 2.85 at $k=0.01$ to 2.74 at $k=0.2$.

Slope estimates are the most accurate across frames, with minimal bias.

$\hat{t}_{dq}$ highly stable, unaffected by k or p . The mean estimate is consistently reliable, similar to other frames.

$\hat{b}_0$ shows moderate increase with k but best performance overall. $p=0.05$ produces lower intercepts than $p=0.01$

$\hat{b}_1$ most stable of all frames (2.74-2.88). Minimal decline with increasing k . $p=0.05$ shows slightly better performance

$\hat{t}_{dq}$ insensitive to parameter changes. Variation <0.05 across all conditions.

Two-Phase Regression Estimation under Multiple Frames when there are response and measurement errors

A subsample of 100 units is taken from the 300 units of the first phase units. The table below provides the totals, biases, and variances of the estimated regression parameters t_{ldq} , b_0 and b_1 under different error conditions. The sample sizes are, $n_1 = 46$, $n_2 = 41$, $n_3 = 42$

Table 5: Two-phase estimation of parameters in frame 1 with varying values of k and p

k	p	$\hat{b}_{10}$	b_{11}	$\hat{t}_{ldq}$	$\hat{b}_{10}$ bias	b_{11} bias	$\hat{b}_{10}$ var	b_{11} var
0.01	0.01	17.0194	2.8036	9290.2340	12.0194	-0.1964	99.5794	0.0274
0.01	0.05	23.1144	2.6926	9251.9446	18.1144	-0.3074	139.1416	0.0384
0.05	0.01	14.6436	2.8473	9274.6324	9.6436	-0.1527	148.6661	0.0412
0.05	0.05	15.9688	2.8210	9286.7174	10.9688	-0.1790	89.1403	0.0246
0.1	0.01	11.9945	2.8862	9254.3532	6.9945	-0.1138	141.0141	0.0391
0.1	0.05	19.4754	2.7573	9340.4799	14.4754	-0.2427	76.7432	0.0208
0.2	0.01	25.7189	2.6465	9271.1282	20.7189	-0.3535	96.5022	0.0264
0.2	0.05	39.8037	2.4216	9262.6310	34.8037	-0.5784	157.8024	0.0436

Table 6: Two-phase estimation of parameters in frame 2 with varying values of k and p

k	P	$\hat{b}_{20}$	$\hat{b}_{21}$	$\hat{t}_{2dq}$	$\hat{b}_{20}$ bias	b_{21} bias	$\hat{b}_{20}$ var	b_{21} var
0.01	0.01	11.2141	2.9025	9212.9009	6.2141	-0.0975	74.1247	0.02080
0.01	0.05	12.4388	2.8853	9170.9089	7.4388	-0.1147	63.0524	0.0179
0.05	0.01	17.0357	2.7932	9165.5941	12.0357	-0.2068	95.0813	0.0268
0.05	0.05	17.1547	2.7893	9178.5246	12.1547	-0.2107	80.1326	0.0225
0.1	0.01	12.7017	2.8653	9180.0763	7.7017	-0.1347	75.4755	0.0212
0.1	0.05	14.9474	2.8298	9267.2687	9.9474	-0.1702	73.4501	0.0202
0.2	0.01	25.4487	2.6492	9176.6328	20.4487	-0.3508	87.2991	0.0245
0.2	0.05	25.8639	2.6659	9258.7329	20.8639	-0.3341	88.0818	0.0246

Table 1: Two-phase estimation of parameters in frame 3 with varying values of k and p

k	P	$\hat{b}_{30}$	$\hat{b}_{31}$	$\hat{t}_{3dq}$	$\hat{b}_{30}$ bias	b_{31} bias	$\hat{b}_{30}$ var	b_{31} var
0.01	0.01	10.3166	2.9222	9236.4608	5.3166	-0.0778	100.3350	0.0281
0.01	0.05	18.3151	2.7832	9183.5607	13.3151	-0.2168	93.4050	0.0264
0.05	0.01	19.2912	2.7562	9144.5399	14.2912	-0.2438	68.4437	0.0194
0.05	0.05	11.1063	2.8907	9192.4186	6.1063	-0.1093	120.5291	0.0337
0.1	0.01	16.5371	2.7981	9247.5162	11.5371	-0.2019	71.4456	0.0197
0.1	0.05	20.3490	2.7332	9138.5297	15.3490	-0.2668	89.8135	0.0254
0.2	0.01	15.5207	2.8262	9204.7988	10.5207	-0.1738	109.6873	0.0308
0.2	0.05	14.1816	2.8421	9195.0037	9.1816	-0.1579	100.4237	0.0281

Frame 1 exhibits the highest variability in bias and variance across parameter combinations. For instance, when $k = 0.2$ and $p = 0.05$, B_0 shows extreme bias (34.80) and variance (157.80), while b_1 has significant negative bias (-0.58) and high variance (0.044). The results suggest that larger k and p values amplify estimation errors, particularly for b_0 . The $\hat{t}_{dq}$ estimates remain stable around 9250–9350, indicating robustness in mean estimation despite parameter fluctuations.

Frame 1 exhibits the most extreme fluctuations in bias and variance, particularly for b_0 . The b_0 bias sharply increases with higher k , peaking at 34.80 when $k = 0.2$, $p = 0.05$, indicating severe overestimation. Similarly, b_0 variance rises significantly at higher k , reaching 157.80, suggesting instability in estimation. The b_1 bias remains negative but worsens with larger k while b_1 variance stays moderate below 0.044.

Compared to Frame 1, Frame 2 demonstrates more moderate bias and variance, though b_0 still shows notable bias, e.g., 20.86 for $k = 0.2$, $p = 0.05$. The variance for b_0 and b_1 is generally lower, with b_1 variance staying below 0.025. The $\hat{t}_{dq}$ estimates are consistent (9150–9267), suggesting reliable mean estimation. The bias patterns imply that higher k worsens b_0 accuracy, while b_1 remains relatively stable.

Frame 2 shows more controlled bias and variance compared to Frame 1. The b_0 bias still increases with k , peaking at 0.86 for $k = 0.2$, $p = 0.05$, but the b_0 variance peaking at 88.08 is significantly lower. The b_1 bias remains stable around -0.35 at worst, and b_1 stays below 0.025, indicating better consistency. This suggests that moderate parameter choices e.g., mid-range k improve stability though higher k still introduces notable bias.

Frame 3 strikes a balance between Frames 1 and 2, with intermediate bias and variance levels. For example, at $k = 0.05$, $p = 0.01$, b_0 has bias 14.29 and variance 68.44, while b_1 shows minimal bias (-0.24) and low variance (0.019). The $\hat{t}_{dq}$ estimates (9138–9247) are consistent, reinforcing stable mean estimation. The results highlight that mid-range k and p values yield better trade-offs between bias and variance.

Frame 3 demonstrates the most balanced performance. The b_0 bias peaking at 15.35 and variance peaking at 120.53 are lower than in Frame 1, while b_1 bias remains mild worst: -0.27 with low variance ≤ 0.034 . Unlike Frame 1, increasing k does not drastically worsen performance, suggesting better robustness across parameter settings.

Frame 1 shows the most extreme variability in parameter estimates. The $\hat{b}_0$ values range widely from 11.99 to 39.80, indicating substantial sensitivity to parameter combinations, particularly at higher k values (0.2). $\hat{b}_1$ remains relatively stable between 2.42–2.89, though shows a clear downward trend with increasing k . The $\hat{t}_{dq}$ estimates are remarkably consistent (9250–9340) despite the volatility in other parameters, suggesting the mean estimation is robust even when slope and intercept estimates fluctuate significantly. The most extreme deviations occur at $k=0.2$, $p=0.05$ where $\hat{b}_0$ spikes to 39.80 while $\hat{b}_1$ drops to 2.42.

Frame 1 exhibits dramatic trends in parameter estimates. The $\hat{b}_0$ shows extreme sensitivity to increasing k values, particularly at $p=0.05$ where it spikes from 18.11 ($k=0.01$) to 39.80 ($k=0.2$). Conversely, $\hat{b}_1$ displays a steady decline from 2.80 to 2.42 as k increases, with the steepest drop occurring at $p=0.05$. The $\hat{t}_{dq}$ remains remarkably stable (9250-9340) despite these fluctuations, forming nearly horizontal lines across all k values. The clear divergence between lines at different p values suggests strong interaction effects between parameters in this frame.

Frame 2 demonstrates more moderate behavior in parameter estimates compared to Frame 1. $\hat{b}_0$ values range from 11.21 to 25.86, showing less extreme variation. $\hat{b}_1$ estimates are tightly clustered between 2.65-2.90, with the lowest values again occurring at $k=0.2$. The $\hat{t}_{dq}$ values (9150-9267) maintain good consistency, similar to Frame 1. This frame shows better stability in estimates overall, though still exhibits the pattern of increasing $\hat{b}_0$ and decreasing $\hat{b}_1$ with higher k values, particularly noticeable at $k=0.2$ where $\hat{b}_0$ exceeds 25.

Frame 2 shows more moderate but still noticeable trends. $\hat{b}_0$ increases linearly from 11.21 to 25.86 as k grows from 0.01 to 0.2, with $p=0.05$ consistently yielding higher estimates than $p=0.01$. $\hat{b}_1$ maintains a gentle downward slope (2.90 to 2.65), with the $p=0.01$ line slightly above $p=0.05$. The $\hat{t}_{dq}$ lines are again nearly flat (9150-9267) and closely clustered, demonstrating excellent stability. The parallel nature of the lines for each p value indicates more predictable behavior than Frame 1, though the systematic changes with k remain evident.

Frame 3 presents the most stable and reliable estimates across all parameters. $\hat{b}_0$ ranges moderately from 10.32 to 20.35, with no extreme outliers. $\hat{b}_1$ maintains a narrow range of 2.73-2.92, showing minimal deviation across parameter combinations. The $\hat{t}_{dq}$ values (9138-9247) are again very consistent. The estimates in this frame show the least sensitivity to changing k and p values, with smooth, predictable variations. Even at $k=0.2$, the maximum $\hat{b}_0$ only reaches 20.35 and $\hat{b}_1$ stays above 2.73, demonstrating superior stability compared to the other frames.

This frame demonstrates the most stable performance. $\hat{b}_0$ shows a controlled increase from 10.32 to 20.35, with $p=0.05$ estimates consistently higher than $p=0.01$. $\hat{b}_1$ exhibits minimal variation (2.92 to 2.73), with the tightest clustering of all frames. The $\hat{t}_{dq}$ lines are again extremely stable (9138-9247) and nearly indistinguishable between p values. The smooth, nearly parallel progression of all lines indicates robust estimation that is less sensitive to parameter changes than other frames, while still maintaining the general pattern of increasing $\hat{b}_0$ and decreasing $\hat{b}_1$ with higher k.

Conclusion

Regression estimation in multiple-frame surveys provides a flexible and efficient method for enhancing survey estimates by integrating auxiliary information across overlapping frames and complex sampling designs. When combined with multiplicity adjustment and two-phase sampling strategies, regression estimators can achieve substantial gains in precision and coherence across frames. However, the presence of measurement and response errors poses serious challenges to the validity of these estimators, as such errors introduce bias and inflate variability, thereby weakening the efficiency gains typically associated with regression-based methods. These effects are often magnified in multiple-frame contexts due to frame overlap, weighting adjustments, and the need to reconcile information from different sources. Evidence from studies on complex survey designs and regression estimation shows that failure to account for these errors can lead to misleading inferences, even when advanced estimation techniques are employed. Consequently, it is essential that measurement and response errors be explicitly incorporated into both the design and analysis stages of multiple-frame surveys, through improved data collection procedures, calibration, model-assisted regression techniques, and adjustments for two-phase sampling. Addressing these error sources enhances the robustness, accuracy, and policy relevance of regression estimates, particularly in official statistics and applied research settings where multiple-frame surveys are increasingly used to achieve comprehensive population coverage.

References

1. Grace YY. Statistical analysis with measurement error or misclassification. New York: Springer-Verlag; 2016.
2. Holsclaw T, Hallgren KA, Steyvers M, Smyth P, Atkins DC. Measurement error and outcome distributions: methodological issues in regression analyses of behavioral coding data. *Psychology of Addictive Behaviors*. 2015;29(4):1031.
3. Kim C, Tamborini CR. Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods and Research*. 2014;43(1):39-72.
4. Legg JC, Fuller WA. Two-phase sampling. In: *Handbook of Statistics*. Vol. 29. Elsevier; 2009. p. 55-70.
5. Lohr S, Rao JK. Estimation in multiple-frame surveys. *Journal of the American Statistical Association*. 2006;101(475):1019-1030.
6. Mecatti F, Singh AC. Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. *Journal de la Société Française de Statistique*. 2014;155(4):51-69.

7. Meijer E, Oczkowski E, Wansbeek T. How measurement error affects inference in linear regression. *Empirical Economics*. 2021;60(1):131-155.
8. Peytchev A. Consequences of survey nonresponse. *The Annals of the American Academy of Political and Social Science*. 2013;645(1):88-111.
9. Preston J. Modified regression estimator for repeated business surveys with changing survey frames. *Survey Methodology*. 2015;41(1):79-98.
10. Rueda MM, Arcos A, Molina D, Ranalli MG. Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. *International Statistical Review*. 2018;86(1):51-67.
11. Rueda MM, Arcos A, Molina D, Ranalli MG. Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. *International Statistical Review*. 2018;86(1):51-67.
12. Rueda M, Cobo B, Perri PF. Randomized response estimation in multiple frame surveys. *International Journal of Computer Mathematics*. 2020;97(1-2):189-206.
13. Singh AC. Generalized multiplicity adjusted multiple frame estimation via regression. In: *Book of Abstracts, Italian Conference on Survey Methodology*. libreriauniversitaria.it; 2013. p. 13.
14. Singh AC, Mecatti F. Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of Official Statistics*. 2011;27(4):633-650.
15. Singh A, Mecatti F. A generalized multiplicity-adjusted Horvitz-Thompson class of multiple frame estimators. In: *Proceedings of the First Italian Conference on Survey Methodology (ITACOSM09)*; 2009.
16. Sitter RR. Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*. 1997;92(438):780-787.