

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2026; 11(4): 108-119
© 2026 Stats & Maths
<https://www.mathsjournal.com>
Received: 08-01-2026
Accepted: 10-02-2026

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

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

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

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

Regression estimation in center sampling 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.2306>

Abstract

In planning and decision-making processes, uncontaminated survey data devoid of measurement and response errors, observation errors and other kind of errors is of essence as this can result in estimates which are biased and which may have inflated standard errors. There is therefore a need to investigate the effects of these errors on survey data when analyzing survey data so as to obtain reliable information for use by statisticians and policy makers. This work looks at the effects of measurement and response errors in center sampling when regression estimation is applied. The work looks into both one phase and two-phase sampling under center sampling. The effect of random errors and biases from systematic errors on simple and correlated response variances under various levels of the errors is carried out. A numerical example is given assuming simple random sampling. The net effect of the errors has been found to inflate simple and correlated response variances and hence overestimation of the variances under different variance estimators. It is therefore recommended that both measurement and response errors be put into consideration when designing and carrying out a survey for more accurate results.

Keywords: Regression estimation, center sampling, measurement error, response error, two-phase sampling, survey data analysis, variance estimation, bias analysis, auxiliary variables, Taylor linearization, sampling techniques, statistical modelling

Introduction

1. Regression estimation in center sampling

The regression model requires adequate information about the auxiliary variable. In the absence of auxiliary variables' information, random selection of a large sample of size n' is carried out where information on the auxiliary variables is obtained, followed by selecting a subsample of size n from the large sample, and observing the study variable y . This sampling process is the double sampling or the two-phase selection process. In this process, randomization is carried out twice, during the first phase and the second phase. In this work, regression estimation is applied in Center Sampling and two-phase Center Sampling.

Let u_{lrq} be the response of the q^{th} unit with profile r in center l . For the whole population, the population mean is

$$\bar{\mu} = \frac{1}{N} \sum_{l=1}^L \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{q=1}^{N_{lr}} u_{lrq}$$

For the l^{th} center, we have

$$\bar{\mu}_l = \frac{1}{N_l} \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{q=1}^{N_{lr}} u_{lrq}$$

The estimator of $\bar{\mu}_l$ is

$$\hat{\mu}_l = \frac{1}{N_l} \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{S_{lr}=1}^{n_{lr}} \frac{u_{lrq}}{\pi_{lrq}}$$

Suppose the variable μ is influenced by auxiliary variables

$x_1, x_2, \dots, x_p$ $x_1, x_2, \dots, x_p$. The finite population values of the r^{th} profile in centre L are assumed to be a realization from an infinite super population model given by $E(\mu_{lrq} / x_{lrq}) = x_{lrq} \beta_{lr}$, $Var(\mu_{lrq}) = \sigma_{lrq}^2$

The estimate of β_{lr} using all the population values of the r^{th} profile in centre L is,

$$\hat{B}_{lr} = \left(\sum_{q=1}^{N_{lr}} x_{lrq} x'_{lrq} \right)^{-1} \left(\sum_{q=1}^{N_{lr}} x_{lrq} \mu_{lrq} \right)$$

$\hat{B}_{lr}$ is now a parameter of the finite population. The objective

will now be of estimating $\hat{B}_{lr}$ and not β_{lr} .

The total is as given by

$$t_{lr} = \sum_{r=1}^{2^L-1} \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \mu_{lrq} \quad \text{which is multiplicity modified total.}$$

Suppose $\mu_{lrq}^0 = \hat{B}_{lr} x_{lrq}$. This is a proxy value of μ_{lrq} . Using the general regression estimator

$$\begin{aligned} \hat{t}_{lr} &= \hat{t}_{\mu_{lr}, \pi} + (t_{x_{lr}} - \hat{t}_{x_{lr}}) \hat{B}_{lr} \quad \text{where} \quad \hat{t}_{\mu_{lr}, \pi} = \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} \\ \hat{t}_{lr} &= \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + \left(\sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} x_{lrq} - \sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{x_{lrq}}{\pi_{lrq}} \right) \hat{B}_{lr} \\ &= \sum_r \frac{1}{m_r} \left(\sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} \right) + \sum_r \frac{1}{m_r} \left(\sum_{r=1}^{N_{lr}} x_{lrq} - \sum_{S_{lr}} \frac{x_{lrq}}{\pi_{lrq}} \right) \hat{B}_{lr} \\ &\quad (1) \\ \hat{B}_{lr} &= \left(\sum_{S_{lr}} \frac{x_{lrq} x'_{lrq}}{\pi_{lrq} \sigma_{lrq}^2} \right)^{-1} \left(\sum_{S_{lr}} \frac{x'_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right) \end{aligned}$$

Where

The above can be written as

$$\begin{aligned} \hat{t}_{lr} &= \sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \sum_{j=1}^p \hat{B}_j x_{jlqr} - \sum_r \frac{1}{m_r} \sum_{S_{lr}} \sum_j \hat{B}_j \frac{x_{jlqr}}{\pi_{lrq}} = \sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + (t_{x_{lr}} - \hat{t}_{x_{lr}})' \hat{B}_{lr} \right] \\ &= \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \hat{\mu}_{lrq} + \sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} - \sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{\hat{\mu}_{lrq}}{\pi_{lrq}} \\ \hat{B}_{lr} &= \hat{T}_{lr}^{-1} \hat{t}_{lr} = \left[\sum_{S_{lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right]^{-1} \left(\sum_{S_{lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right) \end{aligned}$$

$$\begin{aligned} &= \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \hat{\mu}_{lrq} + \sum_r \frac{1}{m_r} \sum_{S_{lr}} \left(\frac{\mu_{lrq} - \hat{\mu}_{lrq}}{\pi_{lrq}} \right) \quad \text{where } \hat{\mu}_{lrq} = x_{lrq} \hat{B}_{lr} \\ &= \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} \hat{\mu}_{lrq} + \sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{e_{lrq}}{\pi_{lrq}} \\ &\quad (2) \end{aligned}$$

Also note that if σ_k^2 is a constant and there is an intercept implying $x_{1k} = 1$ or if σ_k^2 is proportional to one of the x_k s

or if σ_k^2 is proportional to a set of x_k 's, then the adjustment

$$\text{term } e_{ks} = \mu_{lrq} - \hat{\mu}_{lrq} = 0$$

Equation (2) can be written as

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \left[\sum_{q=1}^{N_{lr}} \hat{\mu}_{lrq} + \sum_{S_{lr}} \frac{e_{lrq}}{\pi_{lrq}} \right] = \sum_r \frac{1}{m_r} \left(\sum_{q=1}^{N_{lr}} \hat{\mu}_{lrq} \right)$$

This estimate is not unbiased but approximately unbiased due

to the regression coefficient $\hat{B}_{lr}$ in equation (1) above, note that

$$\sum_q \frac{x_{lrq}}{m_r} = \sum_q \frac{\mu_{lrq}}{m_r} = \begin{bmatrix} x_{1lrq} \\ x_{2lrq} \\ \vdots \\ x_{plrq} \end{bmatrix} = \begin{bmatrix} \sum_q x_{1lrq} \\ \sum_q x_{2lrq} \\ \vdots \\ \sum_q x_{plrq} \end{bmatrix} = \begin{bmatrix} t_{x_{1lr}} \\ t_{x_{2lr}} \\ \vdots \\ t_{x_{plr}} \end{bmatrix} = t_{x_{lr}}$$

Similarly

$$\sum_{S_{lr}} \frac{x_{lrq}}{\pi_{lrq}} = \sum_{S_{lr}} \frac{1}{\pi_{lrq}} \begin{bmatrix} x_{1lrq} \\ x_{2lrq} \\ \vdots \\ x_{plrq} \end{bmatrix} = \begin{bmatrix} \sum_{S_{lr}} \frac{x_{1lrq}}{\pi_{lrq}} \\ \sum_{S_{lr}} \frac{x_{2lrq}}{\pi_{lrq}} \\ \vdots \\ \sum_{S_{lr}} \frac{x_{plrq}}{\pi_{lrq}} \end{bmatrix} = \begin{bmatrix} \hat{t}_{x_{1lr}} \\ \hat{t}_{x_{2lr}} \\ \vdots \\ \hat{t}_{x_{plr}} \end{bmatrix} = \hat{t}_{x_{lr}} \quad (2)$$

We can write,

$$\hat{B}_{lr} = \hat{T}_{lr}^{-1} \hat{t}_{lr} = \left[\sum_{S_{lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right]^{-1} \left(\sum_{S_{lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right)$$

Recall

Thus,

$$\begin{aligned}
\hat{t}_{lr} &= \sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + \left(t_{x_{lr}} - \hat{t}_{x_{lr}} \right) \hat{T}^{-1} \sum_{S_{lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right] \\
&= \sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + \sum_{S_{lr}} \left(t_{x_{lr}} - \hat{t}_{x_{lr}} \right) \hat{T}^{-1} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right] \\
&= \sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \left(1 + \left(t_{x_{lr}} - \hat{t}_{x_{lr}} \right) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lrq}^2} \right) \frac{\mu_{lrq}}{\pi_{lrq}} \right] \\
&= \sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{\mu_{lrq}}{\pi_{lrq}} \quad (3)
\end{aligned}$$

The sample observed value therefore has weights g_{lrqs} and π_{lrq} . While π_{lrq} is given before sampling, g_{lrqs} depends on the sample obtained. Define the quantity E_{lrq} as

$$E_{lrq} = \mu_{lrq} - \tilde{B}_{lr}' x_{lrq} = \mu_{lrq} - \mu_{lrq}^0 \text{ where } \mu_{lrq}^0 = \tilde{B}_{lr}' x_{lrq}.$$

Then we can write

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{(\mu_{lrq}^0 + E_{lrq})}{\pi_{lrq}}$$

But

$$\begin{aligned}
\sum_{S_{lr}} g_{ks} \frac{\mu_{lrq}^0}{\pi_{lrq}} &= \sum_{S_{lr}} \left[1 + \left(t_{x_{lr}} - \hat{t}_{x_{lr}} \right) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lrq}^2} \right] \frac{x_{lrq}' \tilde{B}_{lr}}{\pi_{lrq}} \\
&= \sum_{S_{lr}} \left[\left[\frac{x_{lrq}' \tilde{B}_{lr}}{\pi_{lrq}} \right] + \left(t_{x_{lr}} - \hat{t}_{x_{lr}} \right) \hat{T}^{-1} \hat{T} \tilde{B}_{lr} \right] \\
&= \sum_{S_{lr}} \frac{\mu_{lrq}^0}{\pi_{lrq}} + \sum_{S_{lr}} t_{x_{lr}}' \tilde{B}_{lr} - \sum_{S_{lr}} \hat{t}_{x_{lr}}' \tilde{B}_{lr} \\
\sum_{S_{lr}} g_{ks} \mu_{lrq}^0 &= \sum_{S_{lr}} \frac{\mu_{lrq}^0}{\pi_{lrq}} + t_{lrq}' \tilde{B}_{lr} - \sum_{S_{lr}} \frac{x_{lrq}' \tilde{B}_{lr}}{\pi_{lrq}}
\end{aligned}$$

Thus

$$\begin{aligned}
&= \sum_{S_{lr}} \frac{\mu_{lrq}^0}{\pi_{lrq}} + \sum_{q=1}^{N_{lr}} \mu_{lrq}^0 - \sum_{S_{lr}} \frac{\mu_{lrq}^0}{\pi_{lrq}} \\
&= \sum_{q=1}^{N_{lr}} \mu_{lrq}^0 \text{ which is a constant.}
\end{aligned}$$

Hence

$$\hat{t}_{lr} = \sum_{r=1} \frac{1}{m_r} \left(\sum_{q=1}^{N_{lr}} \mu_{lrq}^0 + \sum_{S_{lr}} g_{lrq} \frac{E_{lrq}}{\pi_{lrq}} \right) \quad (4)$$

For variance estimator, we consider the Taylor linearization approximation of the estimator given by

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} + \left(t_{x_{lr}}' - \hat{t}_{x_{lr}}' \right) \tilde{B}_{lr} \right]$$

$$\begin{aligned}
&= \sum_r \frac{1}{m_r} \left[t_{x_{lr}}' \tilde{B}_{lr} + \sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}} - \sum_{S_{lr}} \frac{x_{lrq}' \tilde{B}_{lr}}{\pi_{lrq}} \right] \\
&= \sum_r \frac{1}{m_r} \left[\sum_{q=1}^{N_{lr}} x_{lrq}' \tilde{B}_{lr} + \sum_{S_{lr}} \frac{\mu_{lrq} - \mu_{lrq}^0}{\pi_{lrq}} \right] \\
&= \sum_r \frac{1}{m_r} \left[\sum_{q=1}^{N_{lr}} \mu_{lrq}^0 + \sum_{S_{lr}} \frac{E_{lrq}}{\pi_{lrq}} \right]
\end{aligned}$$

The approximate variance by Taylor linearization is

$$AV(\hat{t}_{lrq}) = \sum_r \frac{1}{m_r^2} \sum_{q=1}^{N_{lr}} \frac{\Delta_{kk'qq'}}{\pi_{rq'q'}} \frac{E_{lrq}}{\pi_{lrq}} \frac{E_{lrq'}}{\pi_{lrq'}} \quad (5)$$

From equation (4),

$$V(\hat{t}_{lr}) = \sum_r \frac{1}{m_r^2} \sum_{q=1}^{N_{lr}} \sum_{q'=1}^{N_{lr}} x_{lrqq'}' (g_{lrqs} \tilde{E}_{lrq}) (g_{lrq's} E_{lrq's}) \quad (6)$$

But g_{lrqs} is a function of the sample and hence cannot be computed. Using (6) by Taylor linearization approximation, the estimators of the variance can easily be obtained as

$$\hat{V}(\hat{t}_{lr}) = \sum_r \frac{1}{m_r^2} \sum_{S_{lr}} \Delta_{lrqq'} x_{lrqq'}' (g_{lrqs} \tilde{e}_{lrq}) (g_{lrq's} \tilde{e}_{lrq'}) \quad (7)$$

which can be computed from the sample.

From equation (4), another estimation possible is

$$A\hat{V}(\hat{t}_{lr}) = \sum_r \frac{1}{m_r^2} \sum_{S_{lr}} \tilde{\Delta}_{lrqq'} \tilde{e}_{lrq} \tilde{e}_{lrq'} \quad \text{which is equivalent to}$$

putting $g_{lrqs} = 1 \quad \forall q$

2. Two-phase sampling

Assume that the finite population of center L is a realization of a super population generated by the model

$$E_{\lambda}(\mu_{lrq}) = x_{lrq}' \beta_{lr}$$

$$V_{\lambda}(\mu_{lrq}) = \sigma_{lrq}^2$$

The estimate of $\tilde{\beta}_{lr}$ using the realized finite population (size N_{lr} for the r^{th} profile) is

$$\tilde{B}_{lr} = \left(\sum_{U_{lr}} \frac{x_{lrq}' x_{lrq}}{\sigma_{lrq}^2} \right)^{-1} \left(\sum_{U_{lr}} \frac{x_{lrq}}{\sigma_{lrq}^2} \mu_{lrq} \right)$$

$\tilde{B}_{lr}$ is now a parameter of the finite population though an estimate of the model parameter β_{lr} .

Assume a sample S_{1lr} is taken from the finite population having profile r in center L with the inclusion first order and second order inclusion probabilities π_{1lrq} and π_{1lrq^2} respectively. The only observed variables at this point are x_{lrq} for each unit q . Note that had μ_{lrq} been observed, the estimate of $\tilde{B}_{lr}$ would be

$$\hat{\tilde{B}}_{lr} = \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)^{-1} \left(\sum_{S_{1lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)$$

Suppose a second sample S_{lr} a subsample of S_{1lr} is selected where the conditional inclusion probabilities are $\pi_{lrq/S_{1lr}}$ and $\pi_{lrq^2/S_{1lr}}$.

The overall inclusion probability of unit q with profile r in center L is

$$\pi_{lrq} = \pi_{1lrq} \cdot \pi_{lr/S_{1lr}} \text{ and joint inclusion probability is}$$

$$\pi_{lrq^2} = \pi_{1lrq^2} \cdot \pi_{lrq^2/S_{1lr}}$$

Now $\tilde{B}_{1lr}$ can be estimated by

$$\hat{\tilde{B}}_{1lr} = \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)^{-1} \left(\sum_{S_{1lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)$$

Also note that the component $\sum_{S_{1lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}}$ of $\tilde{B}_{1lr}$ can now be estimated as

$$\begin{aligned} \sum_{S_{1lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} &= \sum_{S_{1lr}} \frac{x_{lrq} \hat{\mu}_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} + \sum_{S_{1lr}} \frac{x_{lrq} (\mu_{lrq} - \hat{\mu}_{lrq})}{\sigma_{lrq}^2 \pi_{1lrq} \cdot \pi_{lrq/S_{1lr}}} \\ &= \sum_{S_{1lr}} \frac{x_{lrq} \hat{\mu}_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} + \sum_{S_{1lr}} \frac{x_{lrq} (\mu_{lrq} - \hat{\mu}_{lrq})}{\sigma_{lrq}^2 \pi_{1lrq}} \end{aligned}$$

to get a new

estimate $\tilde{B}_{1lr}^*$ of $\tilde{B}_{1lr}$ as

$$\tilde{B}_{1lr}^* = \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)^{-1} \left(\sum_{S_{1lr}} \frac{x_{lrq} \hat{\mu}_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} + \sum_{S_{1lr}} \frac{x_{lrq} e_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)$$

where $e_{lrq} = \mu_{lrq} - \hat{\mu}_{lrq}$

Note that the estimator $\hat{\tilde{t}}_{lr}$ can be written in the following form

$$\hat{\tilde{t}}_{lr} = \sum_r \frac{1}{m_r} \left(\sum_{S_{1lr}} \frac{\hat{\mu}_{lrq}}{\pi_{1lrq}} + \sum_{S_{1lr}} \frac{\mu_{lrq} - \hat{\mu}_{lrq}}{\pi_{1lrq}} \right)$$

$$= \sum_r \frac{1}{m_r} \left(\sum_{S_{1lr}} \frac{\mu_{lrq}}{\pi_{1lrq}} + \left(\sum_{S_{1lr}} \frac{\hat{\mu}_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{\hat{\mu}_{lrq}}{\pi_{1lrq}} \right) \right) \quad (8)$$

This is easily proved to be unbiased estimate of t_{μ_i} after taking expectation at both stages.

Also

$$\begin{aligned} \hat{\tilde{t}}_{lr} &= \sum_r \frac{1}{m_r} \left(\sum_{S_{1lr}} \frac{\mu_{lrq}}{\pi_{1lrq}} + \left(\sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} \right) \hat{\tilde{B}}_{lr} \right) \\ &= \sum_r \frac{1}{m_r} \left(\sum_{S_{1lr}} \frac{\mu_{lrq}}{\pi_{1lrq}} + \left(\sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} \right) \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)^{-1} \left(\sum_{S_{1lr}} \frac{x_{lrq} \mu_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right) \right) \\ &= \sum_r \frac{1}{m_r} \left(\sum_{S_{1lr}} \left(1 + \left(\sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} \right) \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right)^{-1} \right) \frac{x_{lrq}}{\sigma_{lrq}^2} \frac{\mu_{lrq}}{\pi_{1lrq}} \right) \\ &= \sum_r \frac{1}{m_r} \sum_{S_{1lr}} g_{lrqs} \frac{\mu_{lrq}}{\pi_{1lrq}} \text{ where } g_{lrqs} \text{ is a weight which} \end{aligned}$$

depends on the realized S_{1lr} . The variance is obtained through Taylor linearization where the approximate estimator now is obtained.

Define $\mu_{lrq}^0 = x_{lrq} \tilde{B}_{1lr}$ and let $E_{lrq} = \mu_{lrq} - \mu_{lrq}^0$

$$\hat{\tilde{t}}_{lr} = \sum_r \frac{1}{m_r} \sum_{S_{1lr}} g_{lrqs} \frac{(\mu_{lrq}^0 + E_{lrq})}{\pi_{1lrq}}$$

Then

But

$$\begin{aligned} \sum_{S_{1lr}} g_{lrqs} \frac{\mu_{lrq}^0}{\pi_{1lrq}} &= \sum_{S_{1lr}} \left(1 + \left(\sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} \right) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lrq}^2} \right) \frac{x_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} \\ &= \sum_{S_{1lr}} \frac{x_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} + \left(\sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} \right) T^{-1} \left(\sum_{S_{1lr}} \frac{x_{lrq} x'_{lrq}}{\sigma_{lrq}^2 \pi_{1lrq}} \right) \tilde{B}_{1lr} \\ &= \sum_{S_{1lr}} \frac{x_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} + \sum_{S_{1lr}} \frac{x'_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} - \sum_{S_{1lr}} \frac{x'_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} \\ &= \sum_{S_{1lr}} \frac{x_{lrq} \tilde{B}_{1lr}}{\pi_{1lrq}} = \sum_{S_{1lr}} \frac{\mu_{lrq}^0}{\pi_{1lrq}} \end{aligned}$$

This gives

$$\hat{\tilde{t}}_{lr} = \sum_r \frac{1}{m_r} \left[\sum_{S_{1lr}} \frac{\mu_{lrq}^0}{\pi_{1lrq}} + \sum_{S_{1lr}} \frac{g_{lrqs} E_{lrq}}{\pi_{1lrq}} \right]$$

By Taylor linearization, the approximate estimator is

$$\hat{\tilde{t}}_{lr} \cong \sum_r \frac{1}{m_r} \left[\sum_{S_{1lr}} \frac{\mu_{lrq}^0}{\pi_{1lrq}} + \sum_{S_{1lr}} \frac{E_{lrq}}{\pi_{1lrq}} \right]$$

The approximate variance can now be obtained as

$$\begin{aligned} AV(\hat{t}_{lr}) &= V_1[E_2(\hat{t}_{lr})] + E_1V_2(\hat{t}_{lr}) \\ V_1[E_2(\hat{t}_{\mu_i})] &= V_1\left[\sum_r \frac{1}{m_r} \left[\sum_{S_{lr}} \frac{\mu_{lrq}^0}{\pi_{lrq}} + \sum_{S_{lr}} \frac{E_{lrq}}{\pi_{lrq}} \right]\right] \\ &= V_1\left[\sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{\mu_{lrq}}{\pi_{lrq}}\right] \\ &= \sum_r \frac{1}{m_r^2} \left[\sum_{U_{lr}} \frac{(1-\pi_{lrq})}{\pi_{lrq}} \mu_{lrq}^2 + \sum_{q \neq q'} \sum \frac{\pi_{lrqq'} - \pi_{lrq}\pi_{lrq'}}{\pi_{lrq}\pi_{lrq'}} \mu_{lrq} \mu_{lrq}' \right] \end{aligned}$$

$$\begin{aligned} E_1[V_2(\hat{t}_{\mu_i})] &= E_1V_2\left[\sum_r \frac{1}{m_r} \sum_{S_{lr}} \frac{E_{lrq}}{\pi_{lrq}}\right] \\ &= E_1 \sum_r \frac{1}{m_r^2} V\left[\sum_{S_{lr}} \frac{E_{lrq}}{\pi_{lrq}} I_{lrq} / S_1\right] = E_1 \sum_r \frac{1}{m_r^2} V\left[\sum_{S_{lr}} \frac{E_{lrq}}{\pi_{lrq}} I_{lrq} / S_1\right] \\ AV(\hat{t}_{lr}) &= E_1 \sum_r \frac{1}{m_r^2} \left[\sum_{S_{lr}} \pi_{lrq/s_{lr}} (1-\pi_{lrq/s_{lr}}) \frac{E_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} \sum \pi_{lrqq'} \frac{E_{lrq} E_{lrq'}}{\pi_{lrq} \pi_{lrq'}} \right] \end{aligned}$$

3. Regression estimation in center sampling in the presence of measurement errors

Regression estimation is considered in the presence of measurement and response errors so that we can make use of the information from the auxiliary variable.

Suppose that rather than observing, μ_{lrq} the observed value is y_{lrq} modelled as

$$y_{lrq} = \mu_{lrq} + e_{lrq} + v_{lrq}$$

Where e_{lrq} is a measurement error.

$$\begin{aligned} \text{Let } E_m(y_{lrq}) &= z_{lrq}, \quad V_m(y_{lrq}) = \psi_{lrq}^2 \\ \text{Cov}(y_{lrq}, y_{lrq'}) &= \psi_{lrqq'} \end{aligned} \quad (10)$$

The regression estimator can now be written as

$$\hat{t}_{\mu_i} = \sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{y_{lrq}}{\pi_{lrq}} \quad (11)$$

$$g_{lrqs} = 1 + (t_{x_{lr}} - \hat{t}_{x_{lr}}) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lrq}^2}$$

Where

Due to the presence of measurement errors, consider the process as two-stage to have

$$E(\hat{t}_{\mu_i}) = E_{pm}(\hat{t}_{\mu_i}) = E_p E_m(\hat{t}_{\mu_i})$$

From equation (1),

$$E_p(\hat{t}_{\mu_i}) = \sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{z_{lrq}}{\pi_{lrq}} \quad (12)$$

To obtain expectation under sampling, recall it is a regression estimator which is only approximately unbiased.

From equation (4), we can write

$$E_p E_m(\hat{t}_{\mu_i}) = E_p \left(\sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{z_{lrq}}{\pi_{lrq}} \right) = E_p \sum_r \frac{1}{m_r} \left(\sum_{q=1}^{N_{lr}} z_{lrq}^0 + \sum_{S_{lr}} g_{lrqs} \frac{E_{lrq}}{\pi_{lrq}} \right) \quad (13)$$

Where now $E_{lrq} = z_{lrq} - x_{lrq}' B_{lr}$, where $z_{lrq} = \mu_{lrq} + b_{lrq}$

Using Taylor linearization technique, we have equation (13) as

$$E_p E_m(\hat{t}_{\mu_i}) = E_p \sum_r \frac{1}{m_r} \left(\sum_{q=1}^{N_{lr}} z_{lrq}^0 + \sum_{S_{lr}} E_{lrq} \right) \quad (14)$$

Note that now $g_{lrqs} = \frac{E_{lrq}}{\pi_{lrq}}$ and hence (3.18) can be obtained by linearization

To obtain the variance, consider again the two stages and write

$$V_{p\mu}(\hat{t}_{\mu_i}) = V_p E_{\mu}(\hat{t}_{\mu_i}) + E_p(V_{\mu}(\hat{t}_{\mu_i}))$$

From equation (3), the sampling variance is

$$\begin{aligned} V_p E_{\mu}(\hat{t}_{\mu_i}) &= V_p \left[\sum_r \frac{1}{m_r} \sum_{S_{lr}} g_{lrqs} \frac{z_{lrq}}{\pi_{lrq}} \right] \\ &= \sum_r \frac{1}{m_r^2} \left[\sum_q \pi_{lrq} (1-\pi_{lrq}) \frac{g_{lrqs}^2 z_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} (\pi_{lrqq'} - \pi_{lrq}\pi_{lrq'}) \frac{g_{lrqs} g_{lrq's'} z_{lrq} z_{lrq'}}{\pi_{lrq} \pi_{lrq'}} \right] \\ &= \sum_r \frac{1}{m_r^2} \left[\sum_q \pi_{lrq} (1-\pi_{lrq}) \frac{g_{lrqs}^2 z_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} (\pi_{lrqq'} - \pi_{lrq}\pi_{lrq'}) \frac{g_{lrqs} z_{lrq}}{\pi_{lrq}} \frac{g_{lrq's'} z_{lrq'}}{\pi_{lrq'}} \right] \end{aligned}$$

Also, consider that from equations (4) and (12),

$$V_p E_{\mu}(\hat{t}_{\mu_i}) = V_p \sum_r \frac{1}{m_r} \left[\sum_{q=1}^{N_{lr}} z_{lrq}^0 + \sum_{S_{lr}} g_{lrqs} \frac{E_{lrq}}{\pi_{lrq}} \right]$$

This gives

$$V_p E_{\mu}(\hat{t}_{\mu_i}) = \sum_r \frac{1}{m_r^2} \left[\sum_q \pi_{lrq} (1-\pi_{lrq}) \frac{g_{lrqs}^2 E_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} (\pi_{lrqq'} - \pi_{lrq}\pi_{lrq'}) \frac{g_{lrqs} z_{lrq}}{\pi_{lrq}} \frac{g_{lrq's'} z_{lrq'}}{\pi_{lrq'}} \right]$$

This depends on g_{lrqs} which is a sample weight and hence difficult to compute. However, using the Taylor linearization gives the approximated value as

The response variance is given by

Again, the expectation of equation (16) is complex due to regression effect. A good approximation for large samples is $g_{lrq} = 1$ to get

NB: Usual variance estimators

Using Taylor linearization, we get another estimator with $g_{lrq} = 1$

3.1 Special Cases

a) Simple random sampling for each profile in L

Suppose n_{lr} is the sample size taken from μ_{lr} and simple random sampling without replacement. Then

The estimator is obtained as follows.
First

The response variance term is

The variance estimator is now

$$\frac{1 - \pi_{lrq}}{\pi_{lrq}^2} = \frac{\mu_{lr} (\mu_{lr} - n_{lr})}{n_{lr}^2}$$

From

$$\frac{\pi_{lrqq'} - \pi_{lrq}\pi_{lrq'}}{\pi_{lrq}\pi_{lrq}\pi_{lrq'}} = \frac{-\mu_{lr} (\mu_{lr} - n_{lr})}{n_{lr}^2 (n_{lr} - 1)}$$

the obtained variance estimator is

b) Now consider the special case when $p = 2$, that is assumed super population model is

$$\begin{aligned} \frac{\bar{E}}{V(\mu_{lrq})} \left(\mu_{lrq} \right) &= B_{0lr} + x_{lrq} B_{1lr} \\ V(\mu_{lrq}) &= \sigma_{lr}^2 \end{aligned}$$

Then, $\underline{x}_{lrq} = (1, x_{lrq})$ that is $x_{1lrq} = 1, x_{2lrq} = x_{lrq}$. g_{lrqs} can be obtained as follows

$$g_{lrq} = 1 + \left(\underline{tx}_{lrq} - \hat{\underline{tx}}_{lrq} \right) \hat{T}^{-1} \frac{\underline{x}_{lrq}}{\sigma_{lr}^2}$$

Now

$$\underline{tx}_{lrq} = \sum_q^{N_{lr}} \underline{x}_{lrq} = \sum_q \begin{pmatrix} 1 \\ x_{lrq} \end{pmatrix} = \begin{pmatrix} N_{lr} \\ \sum_q x_{lrq} \end{pmatrix} = \begin{pmatrix} N_{lr} \\ N_{lr} \bar{x}_{lrq} \end{pmatrix}$$

$$\hat{\tilde{t}}_{lrq} = \begin{pmatrix} \hat{t}_{lrq1} \\ \hat{t}_{lrq2} \end{pmatrix} = \begin{pmatrix} \sum_s \frac{1}{\pi_{lrq}} \\ \sum_s \frac{x_{lrq}}{\pi_{lrq}} \end{pmatrix} = \begin{pmatrix} \sum_s \frac{N_{lr}}{n_{lr}} \\ \sum_s \frac{N_{lr} x_{lrq}}{n_{lr}} \end{pmatrix} = \begin{pmatrix} N_{lr} \\ N_{lr} \bar{x}_{lrq} \end{pmatrix}$$

This gives

$$1 + \left(\hat{t}_{lrq} - \tilde{t}_{lrq} \right) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lr}^2} = \left[1 + \begin{pmatrix} 0 \\ N_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) \end{pmatrix} \hat{T}^{-1} \begin{pmatrix} 1 \\ x_{lrq} \end{pmatrix} \frac{1}{\sigma_{lr}^2} \right] \hat{\mu}_i = \sum_r \frac{1}{m_r} \frac{N_{lr}}{n_{lr}} \sum_s \left(1 + \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2} \right) y_{lrq}$$

Obtain $\hat{T}^{-1}$ as follows

$$\begin{aligned} \hat{T} &= \begin{bmatrix} \sum_s \frac{x_{lrq} x'_{lrq}}{\sigma_{lr}^2 \pi_{lrq}} & \sum_s \frac{x_{lrq}}{\sigma_{lr}^2 \pi_{lrq}} \\ \sum_s \frac{x_{lrq}}{\sigma_{lr}^2 \pi_{lrq}} & \sum_s \frac{1}{\sigma_{lr}^2 \pi_{lrq}} \end{bmatrix} = \frac{N_{lr}}{n_{lr} \sigma_{lr}^2} \sum_s \begin{bmatrix} 1 & x_{lrq} \\ x_{lrq} & 1 \end{bmatrix} \\ &= \frac{N_{lr}}{n_{lr} \sigma_{lr}^2} \sum_s \begin{pmatrix} 1 & x_{lrq} \\ x_{lrq} & x_{lrq}^2 \end{pmatrix} = \frac{N_{lr}}{n_{lr} \sigma_{lr}^2} \begin{pmatrix} n_{lr} & \sum_{s_{lr}} x_{lrq} \\ \sum_{s_{lr}} x_{lrq} & \sum_{s_{lr}} x_{lrq}^2 \end{pmatrix} \\ &= \frac{N_{lr}}{n_{lr} \sigma_{lr}^2} \begin{pmatrix} n_{lr} & n_{lr} \bar{x}_{lr} \\ n_{lr} \bar{x}_{lr} & \sum x_{lrq}^2 \end{pmatrix} = \frac{N_{lr}}{\sigma_{lr}^2} \begin{pmatrix} 1 & \bar{x}_{lr} \\ \bar{x}_{lr} & \sum \frac{x_{lrq}^2}{n_{lr}} \end{pmatrix} \end{aligned}$$

$$\hat{T}^{-1} = \frac{\sigma_{lr}^2}{N_{lr}} \frac{1}{\left(\sum \frac{x_{lrq}^2}{n_{lr}} - \bar{x}_{lr}^2 \right)} \begin{pmatrix} \sum \frac{x_{lrq}^2}{n_{lr}} & -\bar{x}_{lr} \\ -\bar{x}_{lr} & 1 \end{pmatrix}$$

$$\begin{aligned} \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lr}^2} &= \frac{1}{N_{lr} \left(\sum \frac{x_{lrq}^2}{n_{lr}} - \bar{x}_{lr}^2 \right)} \begin{pmatrix} \sum \frac{x_{lrq}^2}{n_{lr}} & -\bar{x}_{lr} \\ -\bar{x}_{lr} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_{lrq} \end{pmatrix} \\ &= \frac{1}{N_{lr} \left(\sum \frac{x_{lrq}^2}{n_{lr}} - \bar{x}_{lr}^2 \right)} \begin{pmatrix} \sum \frac{x_{lrq}^2}{n_{lr}} - \bar{x}_{lr} x_{lrq} \\ x_{lrq} - \bar{x}_{lr} \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \left(\hat{t}_{lrq} - \tilde{t}_{lrq} \right) \hat{T}^{-1} \frac{x_{lrq}}{\sigma_{lr}^2} &= \frac{N_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{N_{lr} \left(\sum \frac{x_{lrq}^2}{n_{lr}} - \bar{x}_{lr}^2 \right)} \\ &= \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2} = \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2} \end{aligned}$$

Thus

$$g_{lrqs} = 1 + \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2}$$

Certainly, as $n_{lr} \rightarrow N_{lr}$, $\bar{x}_{lr} \rightarrow \bar{x}_{lr}$ and hence $g_{lrqs} \rightarrow 1$

Note that the estimate of the center total now is

$$\hat{t}_{\mu_i} = \sum_r \frac{1}{m_r} \sum_{s_{lr}} g_{lrqs} \frac{y_{lrq}}{\pi_{lrq}}$$

$$= \sum_r \frac{1}{m_r} \sum_{s_{lr}} \frac{N_{lr}}{n_{lr}} \left(1 + \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2} \right) y_{lrq}$$

That is

$$\left(1 + \frac{n_{lr} (\bar{x}_{lr} - \bar{x}_{lr}) (x_{lrq} - \bar{x}_{lr})}{\sum_{s_{lr}} (x_{lrq} - \bar{x}_{lr})^2} \right) y_{lrq} \quad (18)$$

4. Two phase regression estimation in center sampling in the presence of measurement errors

Let y_{lrq} be the observed value be modelled as

$y_{lrq} = \mu_{lrq} + \xi_{lrq}$ where ξ_{lrq} is an error term. Let

$$V(y_{lrq}) = \psi_{lrq}^2, \text{Cov}(y_{lrq}, y_{lrq'}) = \psi_{lrqq'}.$$

$$\text{Also } E_m(y_{lrq}) = z_{lrq} = \mu_{lrq} + b_{lrq}$$

The estimator now is

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{y_{lrq}}{\pi_{lrq}} + \left(\sum_{s_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{s_{lr}} \frac{x'_{lrq}}{\pi_{lrq}} \right) \hat{B}_{lrq} \right) \quad (19)$$

Also the estimator can be written as

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \sum_{s_{lr}} g_{lrq} \frac{y_{lrq}}{\pi_{lrq}}$$

The estimate now is a result of three phases namely:

- First phase of selecting sample S_{1lr} .
- Second phase of selecting subsample S_{lr} .
- The measurement error phase.

This can be denoted by

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

Using estimator 3.14, we get

$$\begin{aligned} E_m(\hat{t}_{lr}) &= E_m \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{y_{lrq}}{\pi_{lrq}} + \left(\sum_{s_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{s_{lr}} \frac{x'_{lrq}}{\pi_{lrq}} \right) \hat{T}^{-1} \sum_{s_{lr}} \frac{x_{lrq} y_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right) \\ &= \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}}{\pi_{lrq}} + \left(\sum_{s_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{s_{lr}} \frac{x'_{lrq}}{\pi_{lrq}} \right) \hat{T}^{-1} \sum_{s_{lr}} \frac{x_{lrq} z_{lrq}}{\sigma_{lrq}^2 \pi_{lrq}} \right) \\ &= \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}}{\pi_{lrq}} + \left(\sum_{s_{1lr}} \frac{x'_{lrq}}{\pi_{1lrq}} - \sum_{s_{lr}} \frac{x'_{lrq}}{\pi_{lrq}} \right) \hat{B}_{lrq} \right) \end{aligned}$$

Also

$$E_m(\hat{t}_{lr}) = E_\mu \sum_r \frac{1}{m_r} \sum_{s_{lr}} g_{lrq} \frac{y_{lrq}}{\pi_{lrq}}$$

$$= \sum_r \frac{1}{m_r} \sum_{s_{lr}} g_{lrq} \frac{z_{lrq}}{\pi_{lrq}}$$

This can be written as

$$E_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} g_{lrq} \frac{E_{lrqz}}{\pi_{lrqz}} \right) \text{ where,}$$

$$E_{lrqz} = z_{lrq} - x_{lrq} B_{1lrz} = z_{lrq} - z_{lrq}^0 \text{ (Notation of } E_z)$$

Note if we put $E_{lrq} = y_{lrq} - x_{lrq} B_{1lrq} = y_{lrq} - y_{lrq}^0$, then

$$\hat{t}_{lr} = \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{y_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} g_{lrq} \frac{E_{lrq\mu}}{\pi_{lrq}} \right)$$

Under Taylor linearization,

$$\hat{t}_{lr} \cong \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{y_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrq\mu}}{\pi_{lrq}} \right)$$

Taking expectation under measurement errors,

$$E_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \right)$$

$$E_2 E_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \right) = \sum_r \frac{1}{m_r} \sum_{s_{lr}} \frac{z_{lrq}}{\pi_{1lrq}}$$

$$E_1 E_2 E_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r} \sum_{q=1}^{N_{lr}} z_{lrq} = \sum_r \frac{1}{m_r} \left(\sum_{q=1}^{N_{lr}} \mu_{lrq} + \sum_{q=1}^{N_{lr}} b_{lrq} \right) \quad (20)$$

The variance is as below

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

$$V_1 E_2 E_m(\hat{t}_{lr}) = V_1 E_2 \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \right)$$

$$= V_1 \sum_r \frac{1}{\mu_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \right) = V_1 \sum_r \frac{1}{\mu_r} \left(\sum_{s_{lr}} \frac{z_{lrq}}{\pi_{1lrq}} \right)$$

$$= \sum_r \frac{1}{m_r^2} \left[\sum_{q=1}^{N_{lr}} \frac{1 - \pi_{1lrq}}{\pi_{lrq}} z_{lrq}^2 + \sum_{q \neq q'} \pi_{lrqq'} - \pi_{lrq} \pi_{lrq'} \frac{z_{lrq}}{\pi_{lrq}} \frac{z_{lrq'}}{\pi_{lrq'}} \right]$$

$$E_1 V_2 E_m = E_1 V_2 \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{z_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \right)$$

$$= E_1 \sum_r \frac{1}{m_r} \left[\sum_{s_{lr}} \pi_{lrq/s_{lr}} \left(1 - \pi_{lrq/s_{lr}} \right) \frac{E_{lrqz}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} \pi_{lrqq'/s_{lr}} - \pi_{lrq/s_{lr}} \pi_{lrq'/s_{lr}} \frac{E_{lrqz}}{\pi_{lrq}} \frac{E_{lrq'z}}{\pi_{lrq'z}} \right]$$

$$E_1 E_2 V_m = E_1 E_2 V_m \sum_r \frac{1}{m_r} \left(\sum_{s_{lr}} \frac{y_{lrq}^0}{\pi_{1lrq}} + \sum_{s_{lr}} \frac{E_{lrq\mu}}{\pi_{lrq}} \right) = E_1 E_2 \sum_r \frac{1}{m_r^2} \sum_{s_{lr}} \frac{\psi_{lrq}^2}{\pi_{lrq}}$$

Consider

$$E_1 E_2 V_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r} \sum_{s_{lr}} g_{lrq} \frac{y_{lrq}}{\pi_{lrq}}$$

$$= E_1 E_2 \sum_r \frac{1}{m_r^2} \left[\sum_{s_{lr}} g_{lrq}^2 \frac{\psi_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} \sum_{s_{lr}} \frac{g_{lrq} \psi_{lrq}}{\pi_{lrq}} \frac{g_{lrq'} \psi_{lrq'}}{\pi_{lrq'}} \right]$$

For large samples, $g_{lrq} \cong 1$ and

$$E_1 E_2 V_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r^2} \left[\sum_{s_{lr}} \frac{\psi_{lrq}^2}{\pi_{lrq}^2} + \sum_{q \neq q'} \sum_{s_{lr}} \frac{\psi_{lrq}}{\pi_{lrq}} \frac{\psi_{lrq'}}{\pi_{lrq'}} \right] \quad (21)$$

Using simple random sampling without replacement,

$$E_1 E_2 V_m(\hat{t}_{lr}) = \sum_r \frac{1}{m_r^2} \left[\frac{\mu_{lr}^2}{n_{lr}^2} \sum_{s_{lr}} \psi_{lrq}^2 + \frac{\mu_{lr} \mu_{lr}'}{n_{lr} n_{lr}'} \sum_{q \neq q'} \sum_{s_{lr}} \psi_{lrq} \psi_{lrq'} \right] \quad (22)$$

5. Results and Discussion

5.1 Regression estimation in Centre Sampling.

For empirical illustration, simulation of variable X (could be income) and Y (Expenditure) is carried out the objective being to estimate or predict total expenditure which is assumed to regress on income. A population with three centers whose sizes are $N_1 = 118, N_2 = 116, N_3 = 101$ is assumed. With three centers, the number of profiles is eight where

$$r_1 = (1, 0, 0), r_2 = (0, 1, 0), r_3 = (0, 0, 1), r_4 = (1, 1, 0), r_5 = (1, 0, 1), r_6 = (0, 1, 1), r_7 = (1, 1, 1)$$

The model used for generation of the simulated population when there are no measurement errors is for Center 1 is

$$y_{lrq} = B_{lr0} + B_{lr1} x_{lrq} + v_{lrq} = 12 + 0.8 x_{2rq} + v_{2rq} \text{ where}$$

$$v_{2r} \sim N(0, 3) \text{ with}$$

$$E(y_{lrq}) = E(\mu_{lrq}) = B_{lr0} + B_{lr1} x_{lrq} = 5 + 0.2 x_{1rq}$$

The model used for generation of the simulated population when there are no measurement errors is for Center 2 is

$$y_{lrq} = B_{lr0} + B_{lr1} x_{lrq} + v_{lrq} = 15 + 0.3 x_{2rq} + v_{2rq} \text{ where}$$

$$v_{2r} \sim N(0, 2) \text{ with}$$

$$E(y_{lrq}) = E(\mu_{lrq}) = B_{lr0} + B_{lr1} x_{lrq} = 15 + 0.3 x_{2rq}$$

The model used for generation of the simulated population when there are no measurement errors is for Center 3 is

$$y_{lrq} = B_{lr0} + B_{lr1} x_{lrq} + v_{lrq} = 11 + 0.12 x_{3rq} + v_{3rq} \text{ where}$$

$$v_{3r} \square N(0,5) \text{ with}$$

$$E(y_{lrq}) = E(\mu_{lrq}) = B_{lr0} + B_{lr1}x_{lrq} = 11 + 1.2x_{3rq}$$

Introduction of measurement errors on the simulated population is carried out by having the model error term have the distribution say, $y_{lrq} = \mu_{lrq} + e_{lrq} + v_{lrq}$

$$y_{lrq} = \mu_{lrq} + v_{lrq} + e_{lrq}, \text{ where } e_{lrq} \sim N(b_e, \psi^2) \text{ with}$$

$$E(y_{lrq}) = E(\mu_{lrq}) = B_{lr0} + B_{lr1}x_{lrq} + b_e = z_{lrq}.$$

The bias b_e can be put as $b_e = k\mu_{lrq}$ and $\psi^2 = p\mu_{lrq}$, where now constants p and k can be varied for different measurement error levels.

We use simulation to generate data set on the net income of workers to demonstrate linear regression. The data set involves predicting the total expenditure for workers in thousands of dollars (y) given their net income in thousands of dollars (x). This means that for a new net income (x) we will be able to predict the total expenditure (y). A sample of 300 is selected from 500 units, the center sizes are

$$n_1 = 118, n_2 = 116, n_3 = 101$$

The model used is $y_{lri} = B_{lr0} + B_{lr1}x_{lri} + v_{lri}$ where

$$E(v_i) = 0, \quad V(v_i) = \sigma^2.$$

This implies that $E(y_{lri} / x_{lri}) = B_{lr0} + B_{lr1}x_{lri}$. We estimated $B_{lr0} = 5$, $B_{lr1} = 0.3$ and $V(v_{lri}) = 0.5$ so that we can generate the y_{lri} . When there are measurement and response errors, then u is the error term and μ_{lrq} is the true value.

Table 1: Distribution of units of the generated population

Center	Profiles in that Center	Number of Units
Center_1	1, 4, 5, 7	118
Center_2	2, 4, 6, 7	116
Center_3	3, 5, 6, 7	101
Total		335

Table 1 represents the distribution of units of the generated population with the specified regression parameters and population size for each center.

Table 2: Estimation of parameters at Center 1 with varying values of k and p

k	P	$\hat{b}_{10}$	$\hat{b}_{11}$	$\hat{t}_{1rq}$	$b_{10} \text{ bias}$	$b_{11} \text{ bias}$	$b_{10} \text{ var}$	$b_{11} \text{ var}$
0.01	0.01	14.4023	2.8391	21712.8659	9.4023	-0.1609	33.6353	0.0094
0.01	0.05	15.0202	2.8284	21709.2716	10.0202	-0.1716	33.3398	0.0093
0.05	0.01	15.8172	2.8144	21710.9308	10.8172	-0.1856	34.8802	0.0097
0.05	0.05	14.5923	2.8354	21713.7673	9.5923	-0.1646	36.5309	0.0102
0.1	0.01	17.7275	2.7842	21712.5020	12.7275	-0.2158	39.6816	0.0111
0.1	0.05	16.4555	2.8044	21705.7067	11.4555	-0.1956	34.0240	0.0095
0.2	0.01	21.3629	2.7197	21712.4747	16.3629	-0.2803	40.3482	0.0113
0.2	0.05	21.9941	2.7097	21716.3769	16.9941	-0.2903	39.7840	0.0111

Table 3: Estimation of parameters at Center 2 with varying values of k and p

k	P	$\hat{b}_{20}$	$\hat{b}_{21}$	$\hat{t}_{2rq}$	$b_{20} \text{ bias}$	$b_{21} \text{ bias}$	$b_{20} \text{ var}$	$b_{21} \text{ var}$
0.01	0.01	9.7692	2.9223	21406.8490	4.7692	-0.0777	45.3214	0.0126
0.01	0.05	8.9533	2.9357	21403.0952	3.9533	-0.0643	45.6374	0.0127
0.05	0.01	11.2917	2.8966	21406.1774	6.2917	-0.1034	45.2150	0.0126
0.05	0.05	10.3733	2.9125	21401.4171	5.3733	-0.0875	46.0972	0.0129
0.1	0.01	9.5177	2.9261	21403.5057	4.5177	-0.0739	46.7978	0.0131
0.1	0.05	9.3592	2.9264	21397.1806	4.3592	-0.0736	46.7470	0.0130
0.2	0.01	20.6382	2.7426	21406.0180	15.6382	-0.2574	52.0123	0.0145
0.2	0.05	17.8886	2.7878	21411.5397	12.8886	-0.2122	64.0849	0.0179

Table 4: Estimation of parameters at Center 3 with varying values of k and p

k	P	$\hat{b}_{30}$	$\hat{b}_{31}$	$\hat{t}_{3rq}$	$b_{30} \text{ bias}$	$b_{31} \text{ bias}$	$b_{30} \text{ var}$	$b_{31} \text{ var}$
0.01	0.01	12.3516	2.8801	18693.8378	7.3516	-0.1199	44.8076	0.0124
0.01	0.05	12.4045	2.8787	18690.0626	7.4045	-0.1213	47.5897	0.0132
0.05	0.01	13.2816	2.8643	18693.3620	8.2816	-0.1357	46.8087	0.0130
0.05	0.05	13.4327	2.8627	18694.0158	8.4327	-0.1373	45.4496	0.0126
0.1	0.01	11.9928	2.8867	18693.2093	6.9928	-0.1133	44.8339	0.0124
0.1	0.05	14.3216	2.8463	18695.3468	9.3216	-0.1537	46.5653	0.0129
0.2	0.01	17.1960	2.7988	18693.7366	12.1960	-0.2012	54.8035	0.0152
0.2	0.05	17.8902	2.7900	18689.5605	12.8902	-0.2100	50.6657	0.0141

The tables 2, 3, and 4 provide the totals, biases, and variances of the estimated regression parameters b_0 and b_1 under different error conditions.

The bias for b_0 is consistently high across all combinations of k and p , with values ranging from approximately 9.4 to 17.0.

This indicates that the estimated intercept b_0 deviates significantly from the true value, especially for larger values of k and p . The bias for b_1 is smaller but still noticeable, ranging from approximately -0.16 to -0.29. This suggests that the estimated slope b_1 is slightly underestimated compared to the true value. The bias for both b_0 and b_1 tends to increase as k and p increases, indicating that higher values of k lead to less accurate estimates.

The variance for b_0 ranges from approximately 33.6 to 40.3, while the variance for b_1 ranges from approximately 0.009 to 0.011. The variance for both b_0 and b_1 increases slightly as k and p increases, suggesting that higher values of k and p lead to greater uncertainty in the estimates. The variance for b_1 is much smaller compared to b_0 , indicating that the slope estimates are more stable than the intercept estimates.

The bias for b_0 increases as k and p increases, indicating that higher values of k and p lead to larger deviations from the true intercept. The bias for b_1 also increases slightly as k and p increases, suggesting that higher values of k lead to less accurate slope estimates. The effect of p is minimal, as the lines for $p = 0.01$ and $p = 0.05$ are close to each other.

The variance for b_0 increases as k increases, indicating greater uncertainty in the intercept estimates for larger k and p . The variance for b_1 also increases slightly as k increases, suggesting greater variability in the slope estimates for larger k . The effect of p is minimal, as the lines for $p = 0.01$ and $p = 0.05$ are close to each other.

The bias for b_0 ranges from approximately 3.95 to 15.64, which is lower than in Center 1 but still significant. This indicates that the estimated intercept b_0 deviates from the true value, especially for larger k . The bias for b_1 ranges from approximately -0.06 to -0.26, which is smaller than in Center 1. This suggests that the estimated slope b_1 is closer to the true value but still slightly underestimated.

The variance for b_0 ranges from approximately 45.2 to 64.1, while the variance for b_1 ranges from approximately 0.012 to 0.018. The variance for both b_0 and b_1 increases as k and p increases, indicating greater uncertainty in the estimates for larger k and p . The variance for b_1 is much smaller compared to b_0 , similar to Center 1.

The bias for b_0 increases as k and p increases, similar to Center 1, but the bias values are slightly lower. The bias for b_1 increases slightly as k and p increases, but the bias values are smaller than in Center 1.

The variance for b_0 increases as k and p increases, similar to Center 1, but the variance values are slightly higher. The variance for b_1 increases slightly as k and p increases, but the variance values are smaller than in Center 1.

5.2 Regression Estimation in Two Phase Center Sampling

A sub sample of 100 is selected from 300 units of the first phase, where

$$n_1 = 42, n_2 = 45, n_3 = 40$$

The tables 5, 6, and 7 below provides the totals, biases, and variances of the estimated regression parameters b_0 and b_1 under different error conditions.

Table 5: Estimation of parameters at Center 1 with varying values of k and p

k	P	$\hat{b}_{10}$	$\hat{b}_{11}$	$\hat{t}_{1rq}$	b_{10} bias	b_{11} bias	b_{10} var	b_{11} var
0.01	0.01	2.2129	3.0500	9215.5802	-2.7871	0.0500	110.2957	0.0309
0.01	0.05	27.0888	2.6365	9253.5401	22.0888	-0.3635	107.1968	0.0298
0.05	0.01	11.7371	2.8793	9219.3669	6.73707	-0.1207	64.3356	0.0178
0.05	0.05	30.5379	2.5566	9157.9047	25.5379	-0.4434	75.4642	0.0211
0.1	0.01	16.0086	2.8078	9176.3755	11.0086	-0.1922	80.5571	0.0226
0.1	0.05	13.2415	2.8590	9179.9534	8.24151	-0.1410	71.7915	0.0202
0.2	0.01	22.5207	2.7079	9195.2767	17.5207	-0.2921	97.6697	0.0274
0.2	0.05	23.9550	2.6791	9197.5784	18.9550	-0.3209	78.1557	0.0218

Table 6: Estimation of parameters at Center 2 with varying values of k and p

k	P	$\hat{b}_{20}$	$\hat{b}_{21}$	$\hat{t}_{2rq}$	b_{20} bias	b_{21} bias	b_{20} var	b_{21} var
0.01	0.01	5.4295	2.9938	9207.5919	0.4295	-0.0062	67.4317	0.0189
0.01	0.05	16.9801	2.8014	9163.7522	11.9801	-0.1986	119.9984	0.0340
0.05	0.01	11.4679	2.8871	9223.9460	6.4679	-0.1129	72.7607	0.0202
0.05	0.05	-3.9458	3.1573	9248.7566	-8.9458	0.1573	94.4486	0.0263

0.1	0.01	24.6036	2.6689	9197.2806	19.6036	-0.3311	99.9772	0.0280
0.1	0.05	7.35082	2.9566	9291.2359	2.3508	-0.0434	117.6475	0.0322
0.2	0.01	16.0193	2.8199	9161.8551	11.0193	-0.1801	114.7803	0.0325
0.2	0.05	5.03705	3.0109	9237.8415	0.0371	0.0109	163.0511	0.0457

Table 7: Estimation of parameters at Center 3 with varying values of k and p

k	P	$\hat{b}_{30}$	$\hat{b}_{31}$	$\hat{t}_{3rq}$	b_{30} bias	b_{31} bias	b_{30} var	b_{31} var
0.01	0.01	9.7370	2.9224	9200.4847	4.7369	-0.0776	85.5282	0.0240
0.01	0.05	18.8337	2.7724	9263.0700	13.8337	-0.2276	104.9548	0.0291
0.05	0.01	4.9652	3.0010	9222.9354	-0.0348	0.0010	132.7056	0.0370
0.05	0.05	5.9279	2.9930	9281.0308	0.9279	-0.0070	121.0435	0.0335
0.1	0.01	10.7220	2.9103	9253.4382	5.7220	-0.0897	82.8944	0.0230
0.1	0.05	17.1773	2.7991	9276.4823	12.1773	-0.2009	100.2097	0.0276
0.2	0.01	15.7119	2.8245	9298.0389	10.7119	-0.1755	138.8196	0.0381
0.2	0.05	20.9515	2.7413	9294.5856	15.9515	-0.2587	136.4594	0.0376

In tables 5, 6 and 7 bias b_0 shows significant positive bias for higher values of k and p especially for k=0.2 and p=0.05, where the bias is around 18.96.

b_1 has relatively small biases, but they are consistently negative, indicating a slight underestimation of the true parameter.

The variance of b_0 is relatively high, especially for lower k and p values e.g. k=0.01, p=0.01.

The variance of b_1 is generally low, indicating stable estimates across different k and p combinations.

b_0 bias increases significantly as k increases, especially for higher p values e.g., p = 0.05. This indicates that b_0 is consistently overestimated for larger k.

b_1 bias remains relatively small and stable across all k and p values, with a slight negative trend, indicating a consistent underestimation.

The variance of b_0 is high and increases with k, particularly for p = 0.01. This suggests that the estimates of b_0 become less stable as k increases.

The variance of b_1 is low and remains relatively constant across all k and p values, indicating stable estimates.

Bias b_0 exhibits large positive biases for higher k values e.g., k = 0.1, p = 0.01 has a bias of 19.60.

b_1 biases are mostly negative, with the largest bias occurring at k = 0.1, p = 0.01 where bias is -0.33.

The variance of b_0 is high, particularly for k = 0.2, p = 0.05, where it reaches 163.05.

The variance of b_1 is relatively low but increases slightly with higher k and p.

b_0 bias shows a sharp increase for higher k values, especially for p = 0.01. This indicates significant overestimation of b_0 as k grows.

b_1 bias is consistently negative and increases slightly with k, indicating a small but growing underestimation.

The variance of b_0 is very high for larger k values, particularly for p = 0.05, suggesting high instability in b_0 estimates.

The variance of b_1 is low but increases slightly with k, indicating some instability in b_1 estimates for larger k.

6. Conclusion

The presence of measurement, response, observational or other kind of errors in survey data is always a hindrance to appropriate conclusions from the analysis which may lead to wrong decisions in policy making. In this paper, it has shown that the presence of measurement and response errors in center sampling under regression estimation results in overestimating and underestimating the regression parameters. This has occurred both under one phase and two phase sampling process. This therefore calls for identification of error sources and their possible minimization so as to reduce their effects on the data in center sampling. This will improve the efficiency on The planning, forecasting as well policy formulation processes will be greatly improved if measurement and response errors are controlled in the data collection process.

References

1. Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S, editors. *Measurement errors in surveys*. Hoboken: John Wiley & Sons; 2013.
2. Bird TJ, Bates AE, Lefcheck JS, Hill NA, Thomson RJ, Edgar GJ, et al. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*. 2014;173:144-154.
3. Buonaccorsi JP. *Measurement error: models, methods, and applications*. Boca Raton: Chapman and Hall/CRC; 2010.
4. Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*. 2003;98(461):158-168.
5. Fiery MA Jr. A form of two-phase sampling utilizing regression analysis [dissertation]. West Virginia: West Virginia University; 2005.
6. Grajales C. Assumptions of multiple regression: correcting two misconceptions. *Practical Assessment, Research & Evaluation*. 2013;18:1-7.

7. Legg JC, Fuller WA. Two-phase sampling. In: Handbook of statistics. Vol. 29. Amsterdam: Elsevier; 2009. p. 55-70.
8. Mandallaz D, Breschan J, Hill A. New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. Canadian Journal of Forest Research. 2013;43(11):1023-1031.
9. Mecatti F, Migliorati S. Center sampling: a strategy for elusive population surveys. Milan: Department of Statistics, University of Milan-Bicocca; 2003.
10. Sitter RR. Variance estimation for the regression estimator in two-phase sampling. Journal of the American Statistical Association. 1997;92(438):780-787.