https://doi.org/10.56093/JISAS.V80I1.4
Available online at https://isas.live/newsite
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 80(1) 2026 39–42

Estimation of Population Mean in Sample Surveys using Auxiliary Information when it Lacks Temporization

Shalabh and Subhra Sankar Dhar

Indian Institute of Technology Kanpur, Kanpur

Received 25 February 2026; Revised 25 March 2026; Accepted 25 March 2026

SUMMARY

An important feature in estimating the population mean in sample surveys is the use of either current observations or past information, which is available as auxiliary information. The ratio and regression methods of estimation are popular approaches for estimating the population mean using auxiliary information. In most of these surveys, the auxiliary information is available from the past and may change over time when used. This paper examines how this information will affect the estimation process and how the ratio and regression estimation methods are affected by such a change in the auxiliary information.

Keywords: Population mean; Temporization effect; Auxiliary information; Ratio method; Regression method.

1. INTRODUCTION

In survey sampling and statistical estimation, auxiliary information refers to additional data related to the population that is used to improve the accuracy and efficiency of estimators for population parameters such as totals, means, or proportions. The use of auxiliary information to improve the estimation of population parameters assumes that suitable observations on the auxiliary variables are available. Such a supposition is frequently challenged in a variety of practical applications. For example, observations on the auxiliary variables may lack temporization. Methods like ratio estimation, regression estimation, and calibration techniques rely heavily on the availability of such auxiliary variables. The fundamental assumption underlying the use of auxiliary information is that reliable and relevant observations for these variables are available at the time of estimation.

However, in many practical situations, this assumption is difficult to satisfy. One common issue is the lack of temporization, meaning that the auxiliary information does not correspond to the same time period as the study variable. When auxiliary data are outdated, incomplete, or inaccurate, their usefulness in improving statistical estimates is reduced. This discrepancy can introduce bias or reduce the efficiency gains expected from the use of auxiliary variables.

For instance, consider the estimation of the total production of a wheat crop in a particular region. Suppose agricultural statisticians use plot acreage as an auxiliary variable, assuming that crop production is positively correlated with the size of the cultivated area. Ideally, accurate and up-to-date records of the acreage of each plot would be available. In practice, however, such information may come from old land records maintained by local administrative authorities. These records may not reflect recent changes in land ownership, subdivision of plots, changes in cropping patterns, or conversion of agricultural land to non-agricultural use. As a result, the auxiliary variable—plot acreage—may not accurately represent the current situation. When outdated acreage values are used in estimation procedures, the resulting estimate of total wheat production may be less reliable.

Corresponding author: Shalabh
E-mail address: shalab@iitk.ac.in

A similar issue arises in the estimation of total consumption of a food item in the urban areas of a state when districts are treated as sampling units. In such studies, the urban population of each district is a natural auxiliary variable because consumption of most food items is closely related to the number of people living in the area. Typically, the easiest source of this information is past census data, which provides official counts of the urban population. However, census data are collected only at long intervals (often every ten years). Between two censuses, rapid urbanization, migration, or administrative boundary changes can significantly alter the population distribution. Therefore, the census-based population figures used as auxiliary information may not correspond to the actual population at the time of the survey. Because the auxiliary data are not contemporaneous with the study variable, their ability to improve the precision of estimates becomes limited.

In such situations, the auxiliary information is said to be "effaced" or weakened, meaning that its practical value in improving estimation is diminished. While auxiliary variables theoretically enhance estimator performance, their benefits depend critically on timeliness, accuracy, and strong correlation with the study variable. When these conditions are not met, the expected gains from using auxiliary information may not fully materialize.

Another practical example is the estimation of total household electricity consumption in a city, using the number of registered households as an auxiliary variable. Electricity consumption in a region is generally strongly correlated with the number of households, making household count a useful auxiliary variable. In a survey designed to estimate total electricity usage, researchers might rely on municipal housing records or previous census data to obtain the number of households in each ward of the city. However, these records may not be fully up to date. Urban areas often experience rapid construction of new residential buildings, informal housing developments, and population growth driven by migration. If the household counts used as auxiliary information are from municipal records compiled several years earlier, they may significantly underestimate the current number of households. Conversely, some areas might experience depopulation or redevelopment, leading to over estimation. Because the auxiliary variable -household count - does not accurately represent the present conditions, the relationship between the auxiliary variable and electricity consumption becomes distorted. As a result, the estimator that relies on this outdated auxiliary information may produce biased or inefficient estimates of the total electricity consumption of the city.

Similarly, consider the case of knowledge or education, an important auxiliary variable in many social science surveys, such as opinion surveys. Here, the education of a respondent is generally measured by the number of years of schooling. While this is a simple as well as a reasonable measure, it fails to take into account the knowledge cultivated by the respondent with the passage of time, for example, through television programmes, personal experiences, magazines, and newspapers, to namea few. Thus, the observations on the number of years of schooling should be updated. In its absence, the auxiliary information will be somewhat polluted. In such circumstances, a natural question arises: Is it worthwhile to use auxiliary information that lacks temporization?

Auxiliary information can significantly improve statistical estimation, but its effectiveness depends on the availability of accurate, timely, and relevant data. When auxiliary variables are based on outdated records or observations that do not correspond to the current study period, their usefulness diminishes, potentially affecting the reliability of the resulting estimates.

The literature on sampling is rich on the issue of using the auxiliary information; see Cochran (1991), Sukhatme *et al.* (1984), etc., for basic details. It is important to note that this literature lacks a discussion of the temporalization of information on auxiliary variables. Also, the estimation criterion, e.g., uniformly minimum variance unbiased estimators, does not help ensure that such an issue is addressed. Also, goodness-of-fit measures or variable selection criteria have been proposed for different estimation procedures and modelling approaches. For example, the coefficient of determination in multiple linear regression modelling does not address the use of auxiliary information when it lacks temporization; see Shalabh *et al.* (2021, 2025), Dhar *et al.* (2021, 2022, 2026), Shalabh and Dhar (2021, 2023), Dhar and Shalabh (2022), etc. The main objective of this paper is to discuss the problem of temporization, which can occur across many areas and, to the best of our knowledge, remains an unexplored research topic.

This article is an attempt in this direction. To keep the exposition simple and illustrative, we have restricted our attention to the estimation of the population mean using the conventional ratio and regression methods of estimation. The main result is presented in Section 2, followed by some remarks in Section 3.

2. ANALYSIS OF THETEMPORIZATION EFFECT

Let us be given a simple random sample of size n from a finite population of size N for the estimation of the population mean $\bar{Y}$ of the study characteristic Y. The sample mean $\bar{y}$ is well known to be an unbiased estimator of $\bar{Y}$ with variance as

$$V\left(\bar{y}\right)=E\left(\bar{y}-\bar{Y}\right)^{2}=\left(1-\frac{f}{n}\right)S_{y}^{2} \tag{1}$$

where $f=n/N$ is the sampling fraction, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_{i}$,

$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_{i}$, $S_{y}^{2}=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{i}-\bar{Y}\right)^{2}$.

When the information on an auxiliary characteristic X is available, there are several ways to incorporate it to improve the estimation of $\bar{Y}$. Among them, an important way is to use the auxiliary information at the estimation stage. For example, if we employ the conventional ratio and regression methods, we get the following estimators:

$$\bar{y}_{r}=\frac{\bar{y}}{\bar{x}}\bar{X} \tag{2}$$

$$\bar{y}_{lr}=\bar{y}+\frac{s_{xy}}{s_{x}^{2}}\left(\bar{X}-\bar{x}\right) \tag{3}$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_{i}$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_{i},\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_{i}$,

$s_{x}^{2}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}$, $s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)\left(y_{i}-\bar{y}\right)$;

see, e.g., Cochrane (1991), Sukhatme *et al.* (Chapters 5 and 6, 1984).

Both the estimators (2) and (3) are generally biased, and large sample approximations for their mean squared errors are given as

$$MSE(\bar{y}_{r})=\left(\frac{1-f}{n}\right)\left(S_{y}^{2}+R^{2}S_{x}^{2}-2R\rho S_{x}S_{y}\right) \tag{4}$$

$$MSE\left(\bar{y}_{lr}\right)=\left(\frac{1-f}{n}\right)\left(1-\rho^{2}\right)S_{y}^{2} \tag{5}$$

where $R=\left(\frac{\bar{Y}}{\bar{X}}\right)$ and ρ is the population correlation coefficient between X and Y, see Cochrane (1991), Sukhatme *et al.* (Chapters 5 and 6, 1984) for the derivation of these results.

When the observations on the study and auxiliary characteristics in a sample survey are not taken contemporaneously, and the available observations on X are not temporised, the estimators (2) and (3) lose their utility and applicability. Let us therefore suppose that we are given values $x_{1}^{*},x_{2}^{*},\ldots,x_{n}^{*}$, which are not updated due to one reason or another. We further assume that such observations, being far in the past, have increased/decreased in unknown proportions and are expressible as

$$X_{i}^{*}=pX_{i}+U_{i}$$

where p is an unknown constant and U_{i} is a random variable with mean zero, so that $\bar{X}^{*}=p\bar{X}$.

It is assumed that U_{i}'s are stochastically independent of X_{i}'s, X_{i}^{*}'s and Y_{i}'s.

Now, if we employ the effaced information on auxiliary variables, we get the following estimators:

$$\bar{y}_{r}^{*}=\frac{\bar{y}}{\bar{x}^{*}}\bar{X}^{*} \tag{6}$$

$$\bar{y}_{lr}^{*}=\bar{y}+\frac{s_{x^{*}y}}{s_{x^{*}}^{2}}\left(\bar{X}^{*}-\bar{x}^{*}\right). \tag{7}$$

It is easy to verify that both estimators are generally biased. Further, the large sample approximation for their mean squared errors is given by

$$MSE(\bar{y}_{r}^{*})=\left(\frac{1-f}{n}\right)\left[\left(S_{y}^{2}+R^{2}S_{X}^{2}-2R\rho S_{x}S_{y}\right)+\left(\frac{RS_{u}}{p}\right)^{2}\right] \tag{8}$$

$$MSE\left(\bar{y}_{lr}^{*}\right)=\left(\frac{1-f}{n}\right)\left(1-\rho^{2}+\frac{\rho^{2}S_{u}^{2}}{S_{u}^{2}+p^{2}S_{x}^{2}}\right)S_{y}^{2} \tag{9}$$

where $S_{u}^{2}=\frac{1}{N-1}\sum_{i=1}^{N}\left(U_{i}-\bar{U}\right)^{2}$.

Comparing (8) with (4) and (9) with (5), we observe that the use of auxiliary information that is not temporised invariably results in a loss of efficiency, irrespective of the estimation method used.

Next, look at us compare (8) and (9) with (1). It is seen that the ratio method is better than the traditional unbiased method when

$$\rho > \left(\frac{RS_x}{2S_y}\right)\left(1+\frac{S_u^2}{p^2 S_x^2}\right) \tag{10}$$

provided that the quantity on the right-hand side of (10) does not exceed one.

Notice that the range of ρ, as specified by (10), holds for the superiority of the ratio estimator over the sample mean and is shorter than the corresponding range when contemporaneous auxiliary information is available.

On the other hand, discarding auxiliary information is a better strategy when the inequality (10) holds with the opposite sign. In fact, the non-contemporaneous information in the ratio method should not be used at all as long as

$$\left(\frac{RS_x}{2S_y}\right)\left(1+\frac{S_u^2}{p^2 S_x^2}\right) > 1. \tag{11}$$

In the special case of equal coefficients of variation, $\frac{S_y}{\bar{Y}} = \frac{S_x}{X}$, the inequality (11) reduces to $\frac{S_u}{S_x}$ exceeding p.

Interestingly, we observe from (1) in (9) that the regression method continues to provide efficient estimation of the mean relative to the traditional sample mean, an unbiased estimator, whether the auxiliary information is contemporaneous or not.

Finally, from (8) and (9), we observe that the difference between the mean squared errors to the order of our approximation is

$$MSE(\bar{y}_r^*) - MSE\left(\bar{y}_{lr}^*\right)$$
$$= \left(\frac{1-f}{n}\right)\left[\left(RS_x - \rho S_y\right)^2 + \left(\frac{R^2}{p^2} - \frac{\rho^2 S_y^2}{p^2 S_x^2 + S_u^2}\right)S_u^2\right] \tag{12}$$

which is not necessarily non-negative, implying that the ratio method may often outperform the regression method when the auxiliary information is not temporized.

3. SOME REMARKS

We have considered the estimation of the population mean by the conventional ratio and regression methods when contemporaneous observations on the auxiliary characteristics are unavailable and have analysed their efficiencies. The general conclusion emerging from the investigation is that cooperation with this type of auxiliary information should be taken with higher caution.

ACKNOWLEDGEMENTS

The authors are grateful to the referees for their comments, which improved the exposition of the paper.

REFERENCES

Cochran, W.G. (1991). Sampling Techniques, John Wiley.

Shalabh and Subhra Sankar Dhar (2021). "Goodness of Fit in Non-parametric Regression Modelling'', *Journal of Statistical Theory and Practice* (Invited paper for the special issue dedicated to Professor C.R. Rao on "*Celebrating the Centenary of Professor C R Rao*") Journal of Statistical Theory and Practice, volume 15, Article number: 18.

Shalabh, Subhra Sankar Dhar and N. Balakrishna (2021). "Goodness of Fit in Parametric and Non-parametric Econometric Models" in *Optimal Decision Making in Operations Research & Statistics: Methodologies and Applications*, Editors: Leopoldo Eduardo Ca'rdenas-Barro'n, Aquil Ahmed, Irfan Ali and Ali Akbar Shaikh, pp. 68-91, Taylor's & Francis, Routledge, CRC Press.

Shalabh and Subhra Sankar Dhar (2023). "Testing the Goodness of Fit in Instrumental Variables Models" in *G Families of Probability Distributions: Theory and Applications*, Editors: Mir Masoom Al, Irfan Ali, Haitham M. Youso and Mohamed Ibrahim, Taylor's & Francis, CRC Press, pp. 330-343.

Shalabh, Subhra Sankar Dhar and Gaurav Garg (2025). "Robust Measures of Goodness of Fit and Outlier Detection in Linear Regression Models" in *Statistical Outliers and Related Topics*, Editors: Mir Masoom Ali, Irfan Ali, and Haitham M. Yousof, Taylor's & Francis, CRC Press, pp. 462-499.

Subhra Sankar Dhar and Shalabh (2022). "GIVE Statistic for Goodness of Fit in Instrumental Variables Models with Application to COVID Data'' *Nature Scientific Reports*, **12**, 9472.

Subhra Sankar Dhar, Udita Chatterjee and Shalabh (2022). "A Note on Asymptotic Distribution of Trimmed Mean," *Journal of the Indian Society for Probability and Statistics*, **23**, pp. 327-335.

Subhra Sankar Dhar, Shalabh, Prashant Jha, and Aranyak Acharya (2026). "Variable Selection in Multiple Nonparametric Regression Modelling" in *Advanced Mathematical Techniques Applicable in Computational and Intelligent Systems*, Editors: Sandeep Singh, Aliakbar Montazer Haghighi, and Sandeep Dalal, Taylor's & Francis, CRC Press (Accepted for publication).

Sukhatme, P.V., B.V. Sukhatme, S. Sukhatme and C. Asok (1984). Sampling Theory of Surveys with Applications, Indian Society of Agricultural Statistics, New Delhi.