SSC Annual Meeting, June 2025
Proceedings of the Survey Methods Section

Calibration Estimators with Missing Auxiliary Covariates

David Thiessen[1]

ABSTRACT

Various calibration estimators are used in survey sampling, causal inference, data integration, and missing data analysis. These estimators adjust for differences between observed sample characteristics and known population controls. The characteristics could be means or higher moments of variables, functions of observed variables and unknown parameters, or estimated parameters from working regression models. However, these calibration methods assume that reliable population controls are available. In this paper, we consider calibrating a partially observed response variable using partially observed auxiliary variables whose population controls are subject to uncertainty. We propose three estimators that adjust for observed auxiliary covariates, which may not completely align with the observed response variable, and show how the estimators change depending on the precision of the estimated population controls.

KEY WORDS: Calibration, Data Integration, Missing Data, Inverse Probability Weighting.

RÉSUMÉ

Divers estimateurs par calage sont utilisés dans les enquêtes par sondage, l'inférence causale, l'intégration des données et l'analyse des données manquantes. Ces estimateurs corrigent les différences entre les caractéristiques observées de l'échantillon et les contrôles connus de la population. Les caractéristiques peuvent être des moyennes ou des moments supérieurs des variables, des fonctions des variables observées et des paramètres inconnus, ou des paramètres estimés à partir de modéles de régression. Cependant, ces méthodes de calibrage supposent que des contrôles fiables de la population sont disponibles. Dans cet article, nous considérons l'étalonnage d'une variable de réponse partiellement observée à l'aide de variables auxiliaires partiellement observées dont les contrôles de population sont sujets à l'incertitude. Nous proposons trois estimateurs qui s'ajuste aux covariables auxiliaires observées, qui peuvent ne pas correspondre parfaitement à la variable de réponse observée, et montrons comment les estimateurs varient en fonction de la précision des contrôles de population estimés.

MOTS CLÉS : Estimateurs par calage; l'intégration des donnés; données manquantes; pondération inverse des probabilités.

1 INTRODUCTION

Survey sampling calibration is a technique that adjusts weighted sample estimators based on available auxiliary information. The weights are adjusted so that certain weighted sample estimates match externally known values. Calibration provides a framework for using external information and has conceptual appeal. However, calibration generally requires that this auxiliary information is completely available and reliable. In many practical situations, the researcher may be hesitant to use the externally known value as an absolute benchmark because of uncertainty or bias, or the auxiliary information may not be completely available in the sample to use in weighting. In this paper, we propose three different modifications that allow incomplete or uncertain auxiliary information to be used in a calibration framework.

In Section 2, we review the standard calibration estimator. In Section 3, we propose a new method for finding the calibrated weights which allows the external value to be treated as somewhat unreliable. In Section 4, we propose a method that can use a fully observed auxiliary variable in a sample to calibrate a smaller subsample where the response variable is observed. In Section 5, we propose a method for when the auxiliary variable is only available in a subsample but the response variable is fully observed. In Section 6, we provide some closing comments.

[1]MacEwan University, 10700 104 Ave NW, Edmonton, AB, Canada, T5J 4S2, thiessend26@macewan.ca

2 CALIBRATION

Let y_i be a random sample of size n with possibly unequal sampling probabilities π_i. Let $d_i = 1/\pi_i$ be the design weights of the sample. Let T be the unknown population total of Y, which can be unbiasedly estimated by the inverse probability weighting estimator (IPW),

$$\hat{T}_d = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i = \sum_{i=1}^{n} d_i y_i. \tag{1}$$

Suppose the survey also contains data on an auxiliary variable X and that the population total of X, C, is known from a reliable source like a census. A design weighted estimate of C, $\hat{C}_d = \sum d_i x_i$, will likely not be exactly equal to C. Calibration is a technique that tries to find new weights w_i which are close to the design weights but where the new weighted total matches the known population value. Intuitively, if the sample's estimate of C matches the known value of C, that gives some evidence the sample is accurately representing the population. Mathematically, we find new weights w_i which minimize the chi-square distance $\sum(w_i - d_i)^2/d_i$ subject to the calibration constraint that $\sum w_i x_i = C$. Other minimization criteria are available and may be preferred in applications.

After solving this constrained optimization problem with the method of Lagrange multipliers, we use the calibrated weights in an estimator of the total of Y. The well-known calibrated weights and estimator are:

$$\begin{aligned} w_i &= d_i \left(1 + x_i \frac{C - \hat{C}_d}{\sum_{j=1}^{n} d_j x_j^2}\right) \\ \hat{T}_w &= \sum_{i=1}^{n} w_i y_i \\ &= \sum_{i=1}^{n} d_i y_i + \frac{\sum_{i=1}^{n} d_i x_i y_i}{\sum_{i=1}^{n} d_i x_i^2}(C - \hat{C}_d) \\ &= \hat{T}_d + \hat{\beta}_d (C - \hat{C}_d) \end{aligned} \tag{2}$$

Where $\hat{\beta}_d$ is a design weighted estimate of the linear regression coefficient between X and Y. Note that we have not assumed that the true relationship between X and Y is linear. The appearance of a linear regression coefficient is due to the fact that we specified a first-order matching between the known and estimated totals.

3 SKEPTICAL CALIBRATION WITH UNCERTAIN POPULATION CONTROL VALUE

We now introduce some recent developments. Suppose that the total of X, C, is approximately known, but is subject to some uncertainty. For example, the value could be from a census several years ago, from a large study but not a census, or the population under study may be slightly different than the population with the known total. It would still be desirable if the sample estimate matches the population control value, but we suspect the current population's total is not exactly the same as the known population control. Therefore, we won't insist on an exact match between the calibrated estimate and the population control. Instead, we will balance that goal with keeping the calibrated weights closer to the original design weights. Let q_i be the desired "skeptically calibrated" weights.

Mathematically, we introduce a tuning parameter Q which quantifies our level of trust in the given population total. We simultaneously minimize both the differences between the original weights d and the skeptically calibrated weights q, and the difference between the skeptically calibrated auxiliary estimate $\hat{C}_q$ and the population control C. The parameter Q will be a multiplicative factor to the calibration constraint. The larger Q is, the more we trust the population total and the more we emphasize minimizing the difference between the calibrated estimate and the known total. We choose q_i to minimize

$$\sum_{i=1}^{n} \frac{(q_i - d_i)^2}{d_i} + Q\left(C - \sum_{i=1}^{n} q_i x_i\right)^2$$

A population control equality is not required and the minimization problem is easy to solve. The q_i are given by

$$q_i = d_i \left(1 + x_i \frac{Q(C - \hat{C}_d)}{1 + Q \sum d_j x_j^2}\right)$$

Note that as $Q \to 0$, the skeptically calibrated weights go to the original design weights, $q_i \to d_i$ and as $Q \to \infty$, the skeptically calibrated weights go to the ordinary calibration weights, $q_i \to w_i$.

The skeptically calibrated estimate of the population total of Y is

$$\begin{aligned}
\hat{T}_q &= \sum_{i=1}^{n} q_i y_i \\
&= \sum_{i=1}^{n} d_i y_i + \frac{Q \sum_{i=1}^{n} d_i x_i y_i}{1 + Q \sum_{i=1}^{n} d_i x_i^2} \left(C - \hat{C}_d\right) \\
&= \hat{T}_d + \hat{\beta}_Q \left(C - \hat{C}_d\right) \qquad (3)
\end{aligned}$$

As $Q \to 0$, $\hat{T}_q$ goes to the uncalibrated IPW estimator $\hat{T}_q \to \hat{T}_d$. As $Q \to \infty$, $\hat{T}_q$ goes to the ordinary calibration estimator $\hat{T}_q \to \hat{T}_w$.

We also note that $\hat{\beta}_Q$ is a weighted ridge regression estimator with penalty parameter $\lambda = 1/Q$.

$$\begin{aligned}
\hat{\beta}_Q &= \left(1 + Q \sum d_i x_i^2\right)^{-1} \left(Q \sum d_i x_i y_i\right) \\
&= \left(Q^{-1} + S_{dXX}\right)^{-1} S_{dXY}
\end{aligned}$$

Therefore, the skeptically calibrated estimator can be viewed as a shrinkage estimator. When we choose a smaller level of trust Q, the penalty parameter λ increases, $\hat{\beta}_Q$ shrinks towards 0, and the skeptically calibrated estimator (3) shrinks towards the original design-weighted IPW estimate (1).

4 INTERNALLY CALIBRATING SUBSAMPLE OBSERVATIONS TO A MAIN-SAMPLE ESTIMATE

We now consider the situation where there is a completely observed auxiliary variable but the response variable is not fully observed and there is no external estimate of the population total of X. Let (x_i, y_i) be a sample of n individuals where X is available for all individuals but Y is only available for a subset of size m. This may arise because of a subsampling plan to reduce response burden or because of item nonresponse for some individuals. Without loss of generality, let the first m individuals in the sample have Y observed. We will try to use all the available auxiliary information in the main sample, $i = 1, ..., n$, to calibrate the subsample, $i = 1, ..., m$. Intuitively, we are calibrating the subsample to the internal estimate from the entire sample. If the subsample's estimate of the auxiliary total C matches the main-sample's estimate of C, that gives some evidence the subsample is accurately representing the main sample, which should lead to a better estimate of the total of Y.

To expand the previous notation, let π_i be the probability the ith individual is included in the main sample, $d_i = 1/\pi_i$ the corresponding design weights, p_i the probability the ith individual is included in the subsample conditional on being included in the main sample, $s_i = 1/(\pi_i p_i)$ the design weights of the subsample, q_i the internally calibrated weights of the main sample, and a_i the internally calibrated weights of the subsample. We want to find a internally calibrated subsample estimator of the population total of Y, $\hat{T}_a = \sum_{i=1}^{m} a_i y_i$, where the internally calibrated weights q_i and a_i are chosen to minimize

$$\sum_{i=1}^{m} \frac{(a_i - s_i)^2}{s_i} + Q \sum_{i=1}^{n} \frac{(q_i - d_i)^2}{d_i},$$

under the restriction of an in-sample control equation

$$\hat{C}_q := \sum_{i=1}^{n} q_i x_i = \sum_{i=1}^{m} a_i x_i =: \hat{C}_a.$$

Once again Q is a tuning parameter. This time, Q represents the trust we place in the main sample compared to the subsample. The larger we make Q, the more priority we put on keeping the main-sample weights q_i close to their design values d_i, and thus the more we prioritize $\hat{C}_a$ matching $\hat{C}_d$.

The internally calibrated subsample weights are given by

$$a_i = s_i \left(1 + Q x_i \frac{\sum_{j=1}^{n} d_j x_j - \sum_{j=1}^{m} s_j x_j}{\sum_{j=1}^{n} d_j x_j^2 + Q \sum_{j=1}^{m} s_j x_j^2}\right)$$

We again have that as Q goes to 0, the internally calibrated subsample weights go the the design subsample weights, $a_i \to s_i$. As Q goes to infinity, the internally calibrated subsample weights go to ordinary calibrated weights using $\hat{C}_d$ in place of a population control value C.

The internally calibrated subsample estimate of the total is

$$\begin{aligned}\hat{T}_a &= \sum_{i=1}^{m} s_i y_i + \frac{Q \sum_{i=1}^{m} s_i x_i y_i}{\sum_{i=1}^{n} d_i x_i^2 + Q \sum_{i=1}^{m} s_i x_i^2} \left(\sum_{i=1}^{n} d_i x_i - \sum_{i=1}^{m} s_i x_i\right) \\ &= \hat{T}_s + \hat{\beta}_Q \left(\hat{C}_d - \hat{C}_s\right) \end{aligned} \tag{4}$$

As Q goes to 0, the internally calibrated subsample estimator goes to the usual IPW estimator (1) using the design weights from the subsample. As Q goes to infinity, the internally calibrated subsample estimator goes to the ordinary calibrated estimator (2) in the subsample, but with the external population value of C replaced with the design weighted whole-sample estimator $\hat{C}_d$.

β_Q is again a weighted ridge regression estimator of the population regression coefficient. But in this case, the penalty parameter $\lambda = \sum d_i x_i^2 / Q$ depends on both the tuning parameter Q and the design weighted sum of X^2 in the main sample.

5 CALIBRATION WITH INCOMPLETE AUXILIARY VARIABLES

Finally, we consider when the response variable is fully observed but the auxiliary variable could be missing. In a reversal of the above design, assume now that Y is fully observed in a main sample of size n with sampling weights $d_i = 1/\pi_i$ and X is collected from a subsample of size m with combined weights $s_i = 1/\pi_i p_i$. Without loss of generality, X is observed in the first m individuals and missing in the rest. Let C be the known population total of X and T the unknown population total of Y.

Because we know the value of C from some external source, we want to use that information somehow. Our proposal is to use calibration within the subsample where the auxiliary variable X is observed. We will find how calibration changes the subsample design weights $s_i, i = 1, ..., m$. Then, we will apply those same weight modifications to the main sample design weights d_i for individuals $1, ..., m$. Combining the modified weights from individuals $1, ..., m$ with the unmodified weights for individuals $m+1, ..., n$ will give us the desired estimator.

Specifically, we calibrate the subsample estimator $\hat{C}_s = \sum_{i=1}^{m} s_i x_i$ with the control value C. Let w_i be the calibrated weights of the subsample. Using this notation, the calibrated weights are given by

$$w_i = s_i \left(1 + x_i \frac{C - \sum_{j=1}^{m} s_j x_j}{\sum_{j=1}^{m} s_j x_j^2}\right) =: s_i g_i, \quad i = 1, ..., m$$

Let g_i denote the proportional modification that each weight s_i has received. We propose applying this factor to individuals $1, ..., m$ in the main sample, leading to an incomplete calibration estimator

$$\begin{aligned}\tilde{T} &= \sum_{i=1}^{m} d_i g_i y_i + \sum_{i=m+1}^{n} d_i y_i \\ &= \sum_{i=1}^{n} d_i y_i + \left(\frac{\sum_{j=1}^{m} d_j x_j y_j}{\sum_{j=1}^{m} s_j x_j^2} \right) \left(C - \sum_{j=1}^{m} s_j x_j \right) \\ &= \hat{T}_d + \tilde{\beta}(C - \hat{C}_s) \end{aligned} \tag{5}$$

$\tilde{T}$ is similar to the usual calibration estimator (2), except that the regression coefficient $\hat{\beta}_d$ has been replaced with a strange version $\tilde{\beta}$ and the main-sample estimate $\hat{C}_d$ has been replaced with the subsample estimate $\hat{C}_s$. Curiously, $\tilde{\beta}$ involves both the main sample design weights and the subsample design weights. Because the subsample weights s_i are always at least as large as d_i, $\tilde{\beta}$ also seems to have some sort of shrinkage structure compared to the usual $\hat{\beta}$. $\sum_{i=1}^{m} s_j x_j^2$ is an estimate of $\sum_N x_j^2$, but $\sum_{i=1}^{m} d_j x_j y_j$ underestimates $\sum_N x_j y_j$. Therefore, it seems we usually have $|\tilde{\beta}| < |\hat{\beta}_d|$, with the difference become more pronounced the smaller m is. As $m \to 0$, $\tilde{T} \to \hat{T}_d$. As $m \to n$, this incomplete calibration estimator approaches the usual calibration estimator.

6 CONCLUSION

In this paper, we proposed three methods for using unknown or uncertain auxiliary values in calibration. In Section 3, we described a method for when the population control values are not completely reliable. In Section 4, we used the extra available auxiliary information in a sample to calibrate the weights of a selected subsample. In Section 5, we only had auxiliary information available in a subsample but we still used that information to calibrate the portion of the main sample where both variables were observed.

In all three estimators, we saw that they had some shrinkage or regularization aspect. Equations (3) and (4) both resembled ridge regression estimators while equation (5) has some other structure. Intuitively, the more we doubt the auxiliary calibration information, the more the estimator shrinks away from the usual calibration estimator towards the uncalibrated IPW estimate.

Despite their potential, these estimators are still very undeveloped. They have not been studied theoretically or tested in applications. It is also not clear how the tuning parameters Q should be chosen in equations (3) and (4). These estimators will need substantial work to assess their characteristics and whether they could be useful in application.

REFERENCES

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." Journal of the American Statistical Association 87.418: 376-382.

Särndal, Carl-Erik. 2007. "The calibration approach in survey theory and practice." Survey Methodology 33.2: 113-136.

Wu, Changbao. 2023. "Calibration Techniques for Model-Based Prediction and Doubly Robust Estimation." The Survey Statistician Vol. 88: 86-93