Calibration Adjustment for Nonresponse in Cross-Classified Data

Gretchen Falk*

Abstract

In the interest of accurately estimating a parameter of interest, generally a population total, calibration is a method that adjusts the sampling weights of each selected element such that the adjusted estimates of the totals of auxiliary, or benchmark, variables equal the known population totals. Calibration has been used to adjust for frame undercoverage, nonresponse, and sampling weights. To treat nonresponse, under the quasi-randomization model assumptions, the sample of respondents is treated as an additional phase of sampling, where the probabilities of response are estimated from a set of model variables. Under this model and varying response probability assumptions, we explore a special case of the calibration method to treat doubly cross-classified data that uses characteristics of the classification structure as the benchmark and model variables. The resulting calibration estimator can be calculated no matter the minimum sample size over the classification groups and without requiring the collapse of cells, which is its advantage over the poststratified estimator. Theoretical behavior and empirical comparisons of various estimators are presented and discussed.

Key Words: Quasi-randomization, Benchmark, Poststratification

1. Introduction

When population-level information for a finite population is of interest, there are many sampling designs and estimation methods that yield accurate and efficient estimators from which to choose. The unbiasedness and small variance of these estimators depend on complete information from a sample that is carefully chosen to be representative of the population of interest. Nonresponse, particularly if systematic, can drastically worsen the accuracy and efficiency of these estimators by introducing substantial nonresponse bias. This situation commonly arises when the information being collected is sensitive, for instance, financial information.

There are several methods that produce estimators that account for nonresponse and the more familiar of these, including poststratification, involve separating the selected elements into mutually exclusive and homogeneous groups and adjusting, either directly or indirectly, the sampling weights in each group. The poststratified estimator is commonly used, but as the population level estimate combines the group estimates, there are situations in which a poststratified estimate or its variance estimate or both cannot be found or may be unreliable. The poststratification estimator cannot be calculated if any of the poststratification groups have a sample size of zero after the sample has been classified. Additionally, a variance estimate of the poststratification estimator cannot be calculated if any of the groups have a sample size less than two. Therefore, in a situation when the population of interest includes one or more small groups, especially in the cases of small overall sample size and low response rates, it becomes likely that the poststratification estimator will not be a plausible option.

One remedy for this problem is to collapse the groups until each contains enough members for estimation. However, this process can begin to render more difficult interpretation or possibly meaningless interpretation of the components of the estimator. Additionally, the collapse of cells can affect the homogeneity that is assumed to exist in each group, which can yield a less efficient poststratified estimator. We propose a special case of the

*Ernst & Young, 1101 New York Avenue NW, Washington, DC 20005, gretchen.falk@ey.com

calibration method that can be used as an alternative to poststratification because it eliminates the need for collapsing cells. Section 2 will briefly summarize the development of the general calibration method and will present the definitions of the special case proposed here. Section 3 develops and evaluates the behavior of the theoretical model. In Sections 4 and 5, empirical results and further research topics, respectively, will be presented and discussed.

2. Calibration

As Särndal (2007) expresses, calibration is a new name for an existing method of weight adjustment. This method was not originally designed for the treatment of nonresponse, but instead to reduce sampling errors. Considering a sample S of size n_S drawn from the population U of size N with known selection probabilities $\pi_k = 1/d_k$, the standard unbiased estimate of the population parameter T_y is the sum of the product of the usual sampling weights d_k and the value of the variable of interest y_k for each selected element k —

$$\hat{t}_y = \sum_{k \in S} d_k y_k. \tag{1}$$

The calibration method, as presented by Deville and Särndal (1992) among others, instead uses adjusted expansion weights w_k to yield the estimator

$$\hat{t}_{y,cal} = \sum_{k \in S} w_k y_k. \tag{2}$$

These calibration weights w_k are subject the condition that is termed the calibration constraint. For this constraint, we assume that some set of auxiliary variables $\mathbf{x}_k$ with known totals denoted $\mathbf{T_x}$ are available. We then require the estimated totals of the auxiliary variables, using the calibration weights, to be equal to their known totals, and thus the calibration constraint is

$$\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{T_x}. \tag{3}$$

There is a practical advantage to satisfying this constraint in that it guarantees consistency of the estimates of certain benchmark variables over several different surveys conducted by the same or even different organizations. A survey yielding estimates of certain benchmark variables that equal known and accepted totals increases the publicly perceived reliability of the remaining resulting estimates.

Deville and Särndal (1992), who first used the term calibration, also require that some distance function be minimized so that the calibration weights do not largely deviate from the sampling weights. They show that the generalized-regression estimator (GREG) can be expressed in the form of a calibration estimator with weights

$$w_k = d_k(1 + q_k \mathbf{x}_k^T \boldsymbol{\lambda}), \tag{4}$$

where q_k are known positive weights unrelated to the sampling weights and $\boldsymbol{\lambda}$ is the quantity determined by solving the calibration constraint, which can be calculated directly as it is composed of known quantities. Deville and Särndal (1992) then determine that the weights achieved by using several different distance functions are asymptotically equivalent to the generalized-regression estimator. Using the variance expression of the generalized-regression estimator, an estimate of the variance of the calibration estimator can be found. The evaluation of this variance expression suggests that using the calibration weights w_k is advantageous to the sampling weights d_k in terms of estimator variance.

Lundström and Särndal (1999) propose similarly estimated calibration weights to be used to additionally adjust for unit nonresponse under a two-phase model where in the second phase, the respondent sample R of size n_R is drawn from the sample S with known response probabilities $q_k = 1/a_k$. They assume that the adjustment term of the calibration weights in (4) can be used to account for nonresponse bias if all relevant variables are included in the set of benchmark variables. Under this model, the adjustment to the sampling weights can be thought of as an estimate of the response weights a_k.

Following the introduction of this idea by Lundström and Särndal (1999), many modifications were proposed — both to the general sample-level calibration method and specifically to the response-level calibration method — that have yielded more promising models for the unknown response weights a_k. For example, as an alternative to linear calibration weights for nonresponse adjustment, Folsom and Singh (2000) present a class of calibration weights that includes the logistic function as a special case. They also present the idea to use a similar model to adjust for frame undercoverage.

As opposed to minimizing distance functions to determine the set of calibration weights, several authors have chosen to define calibration weights to be some function of the auxiliary variables,

$$w_k = d_k f(\mathbf{x}_k^T \boldsymbol{\lambda}), \tag{5}$$

and to solve for them directly from the calibration constraint equation. From this expression of the calibration weights we can clearly see that, in the case of nonresponse adjustment, the role of this function will be to estimate the response weights a_k. Chang and Kott (2008) term these functions "back-link" functions since, in the context of nonresponse adjustment, these functions are of the form of the inverse or back transformations of link functions found in generalized linear models as discussed in McCullagh and Nelder (1989). Fuller et al (1994) discuss using a functional form of the auxiliary variables to adjust for nonresponse, but define the back-link function to be

$$f(\mathbf{x}_k^T \boldsymbol{\lambda}) = 1 + \mathbf{x}_k^T \boldsymbol{\lambda}, \tag{6}$$

keeping the calibration weights in linear form. Alternatively, Folsom (1991) proposes back-link functions that better reflect the response mechanism, including the logistic and exponential functions. Nonlinear calibration weights are also proposed by Kott (2006).

In all of the calibration estimation schemes proposed prior to the year 2000, only one set of auxiliary variables was being used – the variables with the known population totals to which the adjusted weights were being fit. We will call these variables the benchmark variables. Estavao and Särndal (2000) propose using calibration weights that are estimated using a linear back-link function

$$f(\mathbf{z}_k^T \boldsymbol{\gamma}) = 1 + \mathbf{z}_k^T \boldsymbol{\gamma}, \tag{7}$$

which is dependent on another set of auxiliary variables $\mathbf{z}_k$. This second set of auxiliary variables, that we will call model variables, is required in Estavao and Särndal (2000) to be of the same dimension as the benchmark variables. Kott (2006) also treats the case in which linear calibration weights are estimated using model variables of the same dimension as the benchmark variables. Additionally, he introduces the idea of nonlinear calibration weights depending on these model variables, such as

$$f(\mathbf{z}_k^T \boldsymbol{\gamma}) = e^{\mathbf{z}_k^T \boldsymbol{\gamma}}. \tag{8}$$

Chang and Kott (2008), treating the general nonlinear calibration weights for nonresponse adjustment case, then relax the equal dimension restriction on the model variables and

only require that the set of model variables be of smaller dimension than the benchmark variables. In general, if the number of model variables is less than the number of benchmark variables, the calibration constraint cannot be exactly satisfied for all known control totals. In a manner reminiscent of nonlinear regression, they propose to estimate the response weights and, therefore, the calibration weights by minimizing a quadratic form of the differences between the known benchmark totals and their estimated values.

The calibration estimator that we propose uses benchmark and model variables, both associated differently with the cross-classification structure, and also uses a functional form of the model variables for the definition of the calibration weights w_k. To define the benchmark and model variables, consider a double classification structure in which each element selected to the sample is classified into a particular group based on two chosen characteristics. Assume I row classifications and J column classifications, resulting in $I \times J$ cross-classification groups and $I + J$ marginal groups. To create the special case of calibration proposed here, vectors of membership indicator variables will serve as our definitions of the auxiliary variables. These choices yield an estimator that most accurately models nonresponse while comparing directly to the group-level reweighting method used in poststratification. Additionally, this special case of the calibration method can be used as an alternative to collapsing cross-classification cells for poststratification. As both are reweighting methods that can be used to adjust for nonresponse, many of the authors discussed in this section and also Chang (2012) have presented this idea of using the calibration estimator as an alternative for poststratification. This comparison has been discussed with varying definitions of final sample and auxiliary variables — all of which are different than those presented here.

3. The Special-Case Calibration Model

3.1 Nonresponse Model

For the model developed here, we assume a with-replacement sampling scheme and so will need to modify the usual two-phase model. Consider the following design: In the first phase, choose a with-replacement sample S of size n_S from the population U of size N with probabilities of selection p_k for every element $k \in U$. Under this design, p_k is the probability that element k is chosen on the j^{th} draw or $P(k_j = k) = p_k$ for $j = 1, \ldots, n_S$. Since S is chosen with replacement, let the unadjusted expansion weights be

$$d_{k_j} = \frac{1}{n_S p_{k_j}}$$

as in the *pwr*-estimator discussed in Särndal et al. (1992).

Here we should note that if the sample size is small relative to the population size and if the selection probabilities are small, it is unlikely that any individual will be chosen into the sample more than once. Therefore, while we use the nice theoretical properties of a with-replacement design, the realizations of with- and without-replacement sampling in practice will likely be the same.

In the second phase of this design, a respondent sample is selected from the elements chosen into the sample, S, under a Poisson sampling design. We will assume that any element selected into the sample S more than once will choose to or not to respond to each repeated draw independently. For instance, if element k is selected twice, it would be possible for element k to provide the requested information on one draw and not the other. Realistically, if any element is selected into the sample S more than once, that individual would be given the survey once and the resulting information would be used in the analysis

twice. Adjustments to the model for this more realistic assumption of dependent responses to multiple selections have been made, but will not be discussed here.

Again we should note that if the sample size is small relative to the population size and the selection probabilities are small, then multiple selections of any individual to the sample S is unlikely. Therefore, the realizations of the independent and dependent response mechanisms in practice will be identical. The independent response mechanism assumption used here will allow us to develop needed theoretical properties.

Now, suppose that each element k has a probability of response q_k and let I_{k_j} be an indicator variable for whether the k^{th} element responds to the survey when selected on the j^{th} draw. It follows that $P(I_{k_j} = 1) = q_k$. Note that since I_{k_j}, $j = 1, \ldots, n_S$, are independent, the respondent sample R of size n_R is a with-replacement sample from the population U with probability of selection $p_k q_k$ per draw.

To estimate the parameter of interest, $T_y = \sum_{k \in U} y_k$, define

$$\begin{aligned} \hat{t}_y &= \sum_{i=1}^{n_R} d_{k_i} \frac{1}{q_{k_i}} y_{k_i} \\ &= \frac{1}{n_S} \sum_{j=1}^{n_S} \frac{1}{p_{k_j} q_{k_j}} y_{k_j} I_{k_j}, \end{aligned} \tag{9}$$

which is a mean of independent and identically distributed random variables.

Theorem 1. *The estimator $\hat{t}_y$ in (9) is unbiased under the model and its assumptions presented above.*

The outline of the proof begins by using the usual conditional expectation property,

$$\begin{aligned} E(\hat{t}_y) &= E_S(E_R(\hat{t}_y|S)) \\ &= E_S\Big(E_R\Big(\frac{1}{n_S} \sum_{j=1}^{n_S} \frac{1}{p_{k_j} q_{k_j}} y_{k_j} I_{k_j} \Big| S\Big)\Big). \end{aligned}$$

Remark 1. Recall that I_{k_j} is an indicator variable for whether element k responds when chosen into the sample on the j^{th} draw. Therefore, I_{k_j} is a Bernoulli random variable that is independent and identically distributed with the following properties:

$$E(I_{k_j}) = q_k \tag{10}$$

and

$$\text{Var}(I_{k_j}) = q_k(1 - q_k). \tag{11}$$

Using property (10) in Remark 1,

$$E(\hat{t}_y) = E_S\Big(\frac{1}{n_S} \sum_{j=1}^{n_S} \frac{1}{p_{k_j}} y_{k_j}\Big).$$

Since $W_{k_j} = \frac{1}{p_{k_j}} y_{k_j}$ are independent and identically distributed random variables,

$$\begin{aligned} E(\hat{t}_y) &= \sum_{k \in U} y_k \\ &= T_y. \end{aligned}$$

Since the expected value of $\hat{t}_y$ is the population total of interest, the estimator $\hat{t}_y$ in (9) is unbiased.

Theorem 2. *The estimator $\hat{t}_y$ in (9) has variance*

$$Var(\hat{t}_y) = \frac{1}{n_S}\sum_{k\in U}\frac{1}{p_k q_k}y_k^2 + \frac{1}{n_S}\Big(\sum_{k\in U}p_k - 2\Big)T_y^2 \tag{12}$$

under the model and its assumptions presented in above.

The outline of this proof beings with the usual conditional variance property

$$\text{Var}(\hat{t}_y) \;=\; \text{Var}_S(E_R(\hat{t}_y|S)) + E_S(\text{Var}_R(\hat{t}_y|S)). \tag{13}$$

From the proof of Theorem 1, we know the expression for the inner quantity of the first component, therefore,

$$\text{Var}_S(E_R(\hat{t}_y|S)) \;=\; \text{Var}_S\Big(\frac{1}{n_S}\sum_{j=1}^{n_S}\frac{1}{p_{k_j}}y_{k_j}\Big).$$

We also know that $W_{k_j} = \frac{1}{p_{k_j}}y_{k_j}$, $j = 1,\ldots,n_S$, are independent and identically distributed random variables. Therefore, the first component of (13) is

$$\begin{aligned}\text{Var}_S(E_R(\hat{t}_y|S)) &= \frac{1}{n_S}\sum_{k\in U}\Big(\frac{1}{p_k}y_k - T_y\Big)^2 p_k \\ &= \frac{1}{n_S}\sum_{k\in U}\frac{1}{p_k}y_k^2 - \frac{1}{n_S}\Big(\sum_{k\in U}p_k - 2\Big)T_y^2. \end{aligned}\tag{14}$$

The second component of $\text{Var}(\hat{t}_y)$ in (13) uses property (11) in Remark 1 to yield

$$E_S(\text{Var}_R(\hat{t}_y|S)) \;=\; E_S\Big(\frac{1}{n_S^2}\sum_{j=1}^{n_S}\Big(\frac{1}{q_{k_j}} - 1\Big)\frac{1}{p_{k_j}^2}y_{k_j}^2\Big).$$

Since $W_{k_j} = \Big(\frac{1}{q_{k_j}} - 1\Big)\frac{1}{p_{k_j}^2}y_{k_j}^2$ are independent and identically distributed random variables, the second component of (13) is

$$E_S(\text{Var}_R(\hat{t}_y|S)) \;=\; \frac{1}{n_S}\sum_{k\in U}\Big(\frac{1}{q_k} - 1\Big)\frac{1}{p_k}y_k^2. \tag{15}$$

Combining the two components – (14) and (15) – of $\text{Var}(\hat{t}_y)$ yields the variance expression in (12).

Here it is interesting to note that the variance includes two components, both of which depend on the selection probabilities. However, only one of the components depends on the response probabilities.

3.2 The Nonresponse Model with Unknown Response Probabilities

In Section 3.1 we assumed that the probabilities of response q_k were known, which is generally an unreasonable assumption. In order to use estimated probabilities of response, some adjustments must be made to the model. Now, assume that the true response probabilities can be modeled as

$$q_k = \frac{1}{f(\mathbf{z}_k^T\boldsymbol{\beta}_0)}, \tag{16}$$

where $f(\eta)$ is a monotonic and twice-differentiable function with first derivative denoted as $f_1(\eta)$, where $\mathbf{z}_k$ denotes the model variables, and where $\boldsymbol{\beta}_0$ is the true value of an unknown response parameter, $\boldsymbol{\beta}$.

In order to estimate the parameter of interest $T_y = \sum_{k \in U} y_k$, let the calibration estimator be

$$\begin{aligned} \hat{t}_{y,cal} &= \sum_{i=1}^{n_R} d_{k_i} f(\mathbf{z}_{k_i}^T \hat{\boldsymbol{\beta}}) y_{k_i} \\ &= \frac{1}{n_S} \sum_{j=1}^{n_S} \frac{1}{p_{k_j}} f(\mathbf{z}_{k_j}^T \hat{\boldsymbol{\beta}}) y_{k_j} I_{k_j}, \end{aligned} \tag{17}$$

where $\hat{\boldsymbol{\beta}}$ estimates the response parameter by minimizing the optimization function

$$R(\boldsymbol{\beta}) = (\mathbf{T_x} - \hat{\mathbf{t}}_{\mathbf{x}}(\boldsymbol{\beta}))^T (\mathbf{T_x} - \hat{\mathbf{t}}_{\mathbf{x}}(\boldsymbol{\beta})), \tag{18}$$

where $\mathbf{T_x}$ are the known totals of the benchmark variables, which are denoted by $\mathbf{x}_k$, and where

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{x}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n_R} d_{k_i} f(\mathbf{z}_{k_i}^T \boldsymbol{\beta}) \mathbf{x}_{k_i} \\ &= \frac{1}{n_S} \sum_{j=1}^{n_S} \frac{1}{p_{k_j}} f(\mathbf{z}_{k_j}^T \boldsymbol{\beta}) \mathbf{x}_{k_j} I_{k_j}. \end{aligned} \tag{19}$$

Note that if $\boldsymbol{\beta}_0$ is known, then $f(\mathbf{z}_{k_j}^T \hat{\boldsymbol{\beta}}) = f(\mathbf{z}_{k_j}^T \boldsymbol{\beta}_0) = \frac{1}{q_{k_j}}$ and this estimator simplifies to the estimator $\hat{t}_y$ in (9) in Section 3.1.

3.2.1 The Two-Way Cross-Classification Definitions for Calibration Estimation

Recall the two-way cross-classification structure presented in Section 3. For our special case of the calibration method to compare with poststratification, we will define the benchmark variables, $\mathbf{x}_k$, as a P-vector of indicator variables of cross-classification cell membership, where $P = I \times J$. Therefore, $\mathbf{T_x}$ will be a P-vector of cross-classification cell population totals.

We will define the model variables, $\mathbf{z}_k$, to be a Q-vector of indicator variables of the marginal classifications for element k. Therefore, each $\mathbf{z}_k$ will be a vector of dimension $Q \leq P$ with a 1 in two places. For ease of interpretation, define $Q = I + J$ with one indicator variable for each of the I levels and one indicator variable for each of the J levels. However, one should note that using an $I + J$ vector will result in singularity while solving the estimating equation (18) since one degree of freedom is lost due to the fact that both sets of marginal totals will equal the population total. Therefore, Q should be defined by $I + J - 1$ for calculation.

By defining the benchmark and model variables in this way, we are making the assumption that the probabilities of response are equal for all elements k belonging to the same cross-classification group. Also, we are requiring knowledge of the cross-classification structure of the population with our definition of $\mathbf{T_x}$.

Now, if we let

$$\theta_g = \frac{1}{N_g} \sum_{k \in U_g} y_k \tag{20}$$

be the mean of the variable of interest in the g^{th} cross-classification group and $\boldsymbol{\theta}$ be a P-vector of the population means θ_g, then another characteristic of two-way cross-classification is that

$$\begin{aligned} y_k &= \mathbf{x}_k^T\boldsymbol{\theta} + \epsilon_k \\ &= \boldsymbol{\theta}_g + \epsilon_k \end{aligned} \tag{21}$$

with the error term ϵ_k following the condition

$$\sum_{k\in U_g} \epsilon_k = 0. \tag{22}$$

3.2.2 The Consistency and Distribution of $\hat{\boldsymbol{\beta}}$

In order to examine the behavior of the calibration estimator in (17) , we will need to employ methods that require $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O\Big(\frac{1}{\sqrt{n_S}}\Big)$ to hold. The proofs for the following theorems are lengthy and not shown here.

Theorem 3. *Under the model and its assumptions defined in Section 3.2, $R_{n_S}(\boldsymbol{\beta}) \to R_\infty(\boldsymbol{\beta})$ almost surely.*

Assumption 1. $\boldsymbol{\beta}_0$ is the unique minimum of $R_\infty(\boldsymbol{\beta})$.

Theorem 4. *In any compact set which contains $\boldsymbol{\beta}_0$ as an interior point, $R_{n_S}(\boldsymbol{\beta}) \to R_\infty(\boldsymbol{\beta})$ uniformly in probability.*

Theorem 5. *In any neighborhood of $\boldsymbol{\beta}_0$ defined as in Assumption 1, $\hat{\boldsymbol{\beta}}_{n_S} \to \boldsymbol{\beta}_0$ in probability.*

Theorem 6. *$\sqrt{n_S}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges in distribution to a Normal distribution with mean zero and constant variance $\mathbf{V}_{\boldsymbol{\beta}}$.*

After several lengthy derivations not shown here that make use of the Taylor Series of $g(\hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$, we find that $\hat{\mu}_{y,cal} = 1/N\hat{t}_{y,cal}$ can be approximated by

$$\hat{\mu}_{y,cal} = \hat{\mu}_y + (\boldsymbol{\mu}_{\mathbf{x}} - \hat{\boldsymbol{\mu}}_{\mathbf{x}})^T\mathbf{P}\boldsymbol{\theta} + O_p\Big(\frac{1}{n_S}\Big), \tag{23}$$

where

$$\hat{\mu}_y = \frac{1}{N}\hat{t}_y, \tag{24}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}} &= \frac{1}{N}\mathbf{T}_{\mathbf{x}} \\ &= \frac{1}{N}\sum_{k\in U}\mathbf{x}_k \\ &= \frac{1}{Nn_S}\sum_{j=1}^{n_S}\frac{1}{p_{k_j}}f(\mathbf{z}_{k_j}^T\hat{\boldsymbol{\beta}})\mathbf{x}_{k_j}I_{k_j}, \end{aligned}$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \frac{1}{Nn_S}\sum_{j=1}^{n_S}\frac{1}{p_{k_j}}f(\mathbf{z}_{k_j}^T\boldsymbol{\beta}_0)\mathbf{x}_{k_j}I_{k_j},$$

$$\mathbf{P} = \boldsymbol{\mu}_{\mathbf{xz}}(\boldsymbol{\mu}_{\mathbf{xz}}^T\boldsymbol{\mu}_{\mathbf{xz}})^{-1}\boldsymbol{\mu}_{\mathbf{xz}}^T, \tag{25}$$

$$\boldsymbol{\mu}_{\mathbf{xz}} = \frac{1}{N}\sum_{k\in U}\frac{f_1(\mathbf{z}_k^T\boldsymbol{\beta}_0)}{f(\mathbf{z}_k^T\boldsymbol{\beta}_0)}\mathbf{x}_k\mathbf{z}_k^T. \tag{26}$$

This approximation is useful for the evaluation of the theoretical properties of the calibration estimator because it only depends on the true response parameter $\boldsymbol{\beta}_0$ and not its estimated value $\hat{\boldsymbol{\beta}}$, which varies from sample to sample.

Theorem 7. *The estimator $\hat{\mu}_{y,cal}$ in (23), under the unknown response probability model and its assumptions is unbiased to the order $O(\frac{1}{n_S})$.*

The outline of this proof starts with the expected value of the useful approximation in (23)

$$E(\hat{\mu}_{y,cal}) \quad = \quad E(\hat{\mu}_y) + (\boldsymbol{\mu}_{\mathbf{x}} - E(\hat{\boldsymbol{\mu}}_{\mathbf{x}}))^T\mathbf{P}\boldsymbol{\theta} + O\Big(\frac{1}{n_S}\Big). \tag{27}$$

From Theorem 1 we know that

$$E(\hat{t}_y) = T_y$$

thus

$$E(\hat{\mu}_y) \quad = \quad \mu_y. \tag{28}$$

As $\hat{\mathbf{t}}_{\mathbf{x}}$ is of the same form as $\hat{t}_y$ but with y_{k_j} substituted by $\mathbf{x}_{k_j}$, it also follows that

$$E(\hat{\boldsymbol{\mu}}_{\mathbf{x}}) \quad = \quad \boldsymbol{\mu}_{\mathbf{x}}. \tag{29}$$

Substituting the expected values into (27), yields

$$E(\hat{\mu}_{y,cal}) = \mu_y + O\Big(\frac{1}{n_S}\Big). \tag{30}$$

Therefore, $\hat{\mu}_{y,cal}$ is unbiased to the order of $O\Big(\frac{1}{n_S}\Big)$.

From this result we determine that calibration adjustment leads to an accurate estimate to the order $O\Big(\frac{1}{n_S}\Big)$ in the presence in nonresponse.

Theorem 8. *The estimator $\hat{\mu}_{y,cal}$ in (23), under the unknown response probability model and its assumptions, has variance*

$$\mathit{Var}(\hat{\mu}_{y,cal}) = \mathit{Var}(\hat{\mu}_y) - \boldsymbol{\theta}^T\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\theta} + \boldsymbol{\theta}^T(\mathbf{I}-\mathbf{P})^T\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I}-\mathbf{P})\boldsymbol{\theta} \tag{31}$$

to the order $O\Big(\frac{1}{n_S^{3/2}}\Big)$, where $\mathit{Var}(\hat{\mu}_y)$ is $\frac{1}{N^2}\mathit{Var}(\hat{t}_y)$ — found in Theorem 2, $\mathbf{I}$ is a $P \times P$ identity matrix, and

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \mathit{Var}(\hat{\boldsymbol{\mu}}_{\mathbf{x}}). \tag{32}$$

The proof is lengthy and not shown here.

4. Results and Discussion

4.1 Introduction

In this section we look at some empirical results of the special-case calibration method presented in Section 3.2. To do so, we use four artificial populations, each based upon characteristics of a set of data from the Quarterly Census of Employment and Wages conducted in the first quarter of 2005 by the Bureau of Labor Statistics. Each of the $N = 283,725$ businesses is classified by the state in which they are located — A, B, C, D, or E — and by the type of industry to which they belong — 1, 2, 3, 4, or 5. The variable of interest is total quarterly wages. Each of the four populations is based on the parameters given in Table 1 with F_{ij} representing the percentage of the population that belongs to the ij^{th} cross-classification group.

Table 1: Population Parameters

	1	2	3	4	5
A	$N_{11} = 5986$ $F_{11} = 2.11$ $\theta_{11} = 2,991,523$	$N_{12} = 5548$ $F_{12} = 1.96$ $\theta_{12} = 3,854,097$	$N_{13} = 7712$ $F_{13} = 2.72$ $\theta_{13} = 5,812,704$	$N_{14} = 3969$ $F_{14} = 1.40$ $\theta_{14} = 17,760,295$	$N_{15} = 1299$ $F_{15} = 0.46$ $\theta_{15} = 4,158,368$
B	$N_{21} = 18,782$ $F_{21} = 6.62$ $\theta_{21} = 1,048,093$	$N_{22} = 31,572$ $F_{22} = 11.1$ $\theta_{22} = 1,228,337$	$N_{23} = 22,012$ $F_{23} = 7.76$ $\theta_{23} = 4,630,796$	$N_{24} = 4982$ $f_{24} = 1.76$ $\theta_{24} = 5,122,252$	$N_{25} = 4504$ $F_{25} = 1.59$ $\theta_{25} = 730,731$
C	$N_{31} = 13,518$ $F_{31} = 4.76$ $\theta_{31} = 1,293,414$	$N_{32} = 13,099$ $F_{32} = 4.62$ $\theta_{32} = 1,706,660$	$N_{33} = 17,837$ $F_{33} = 6.29$ $\theta_{33} = 4,112,411$	$N_{34} = 5610$ $F_{34} = 1.98$ $\theta_{34} = 7,687,645$	$N_{35} = 3001$ $F_{35} = 1.06$ $\theta_{35} = 1,761,251$
D	$N_{41} = 30,428$ $F_{41} = 10.7$ $\theta_{41} = 708,971$	$N_{42} = 36,017$ $F_{42} = 12.7$ $\theta_{42} = 758,204$	$N_{43} = 32,541$ $F_{43} = 11.5$ $\theta_{43} = 2,104,408$	$N_{44} = 10,963$ $F_{44} = 3.86$ $\theta_{44} = 4,273,129$	$N_{45} = 5399$ $F_{45} = 1.90$ $\theta_{45} = 640,548$
E	$N_{51} = 2225$ $F_{51} = 0.78$ $\theta_{51} = 7,418,207$	$N_{52} = 2020$ $F_{52} = 0.71$ $\theta_{52} = 10,368,820$	$N_{53} = 3110$ $F_{53} = 1.10$ $\theta_{53} = 21,441,100$	$N_{54} = 1076$ $F_{54} = 0.38$ $\theta_{54} = 44,797,328$	$N_{55} = 515$ $F_{55} = 0.18$ $\theta_{55} = 11,421,101$

Each of the four populations was created such that the individual values of total quarterly wage, y_k, for the members in each cross-classification group follow a Normal distribution with the means in Table 1 and varying values of standard deviation. The sets of standard deviations used to create Populations 1 through 4, respectively, were constant for all groups, varying and proportional to the mean in each group, constant for each state but varying for each industry group, and simply varying for each group. Specifically, the standard deviation used for all cross-classification groups to create Population 1 is proportional to the smallest group mean $\theta_{45} = 640,548$ and, therefore, small relative to the larger group means. These four populations yielded the following totals of the quarterly wages: $T_{y,1} = 8,366,800$, $T_{y,2} = 8,363,461$, $T_{y,3} = 8,356,790$, and $T_{y,4} = 9,061,540$, all in hundred of thousands of dollars.

Once the populations were created, a set of 10,000 samples was taken from each population for each three sample sizes — $n_S = 500$, $n_S = 2000$, and $n_S = 5000$ — using simple random sampling. Next, a respondent sample was selected from each sample using each of four different true response parameters, yielding average response rates of approx-

imately 30%, 50%, 70%, and 80%, in the inverse response function $f(\eta) = 1 + \exp(-\eta)$.

Using each sample in each of these several sets of 10,000 respondent samples, the calibration estimator in (17) and the poststratification estimator

$$\hat{t}_{y,p} = \sum_{g \in G} \frac{N_g}{n_{R_g}} \sum_{i \in R_g} y_i$$

were calculated.

From each set of 10,000 estimates, the bias, variance, standard deviation, mean squared error, and root mean squared error were calculated. Using these quantities, estimator efficiencies were found using both mean squared error and variance to compare special-case calibration and poststratification. Additionally, a tally was kept for each method of the number of samples that did not iteratively converge, as in calibration, or that could or should not be estimated, as in poststratification. The cases where the poststratified estimate could not be calculated were not included in the determination of the bias and variance.

4.2 The Comparison of Calibration and Poststratification

Because of the real potential that many of the samples when beginning with $n_S = 500$ and $n_S = 2000$ would have at least one cross-classification group with no responding individuals, the poststratified estimator was only calculated for the case when $n_S = 5000$. Table 2 shows the relative efficiencies of the special-case calibration estimator to the poststratified estimator with respect to both mean squared error and variance.

Table 2: Relative Efficiencies of Special-Case Calibration to Poststratification

	Pop 1	Pop 2	Pop 3	Pop 4
30%	$e_{MSE} = 0.0002$	$e_{MSE} = 0.092$	$e_{MSE} = 0.094$	$e_{MSE} = 0.061$
	$e_{Var} = 0.002$	$e_{Var} = 0.637$	$e_{Var} = 0.667$	$e_{Var} = 0.474$
50%	$e_{MSE} = 0.003$	$e_{MSE} = 0.746$	$e_{MSE} = 0.735$	$e_{MSE} = 0.621$
	$e_{Var} = 0.004$	$e_{Var} = 0.826$	$e_{Var} = 0.820$	$e_{Var} = 0.704$
70%	$e_{MSE} = 0.00004$	$e_{MSE} = 0.018$	$e_{MSE} = 0.017$	$e_{MSE} = 0.011$
	$e_{Var} = 0.003$	$e_{Var} = 0.426$	$e_{Var} = 0.420$	$e_{Var} = 0.360$
80%	$e_{MSE} = 0.0001$	$e_{MSE} = 0.068$	$e_{MSE} = 0.067$	$e_{MSE} = 0.044$
	$e_{Var} = 0.0001$	$e_{Var} = 0.079$	$e_{Var} = 0.078$	$e_{Var} = 0.052$

As we can see from Table 2, the calibration estimator has a higher variance and a much higher bias than the poststratified estimator in Population 1. Recall that Population 1 was created using the same error variance in all cross-classification groups. In fact, the standard deviation used was proportional to the smallest group mean $\theta_{45} = 640,548$ and, therefore, relatively small compared to the larger group means. Therefore, the values of the total quarterly wage, y_k, for businesses in the same cross-classification group with a larger mean are generally homogeneous. On the other hand, when the standard deviation of the variable of interest is proportional to the size of the mean in each group and, therefore, is larger for these groups with larger means, then all of the cross-classification groups are less homogeneous with respect to total quarterly wage. The poststratification method is most effective when the poststratification groups are as homogeneous as possible, which

explains the better performance poststratification shows over special-case calibration in Population 1. In fact, over the different populations, the calibration estimator performs fairly consistently and so the changes in the relative efficiencies that we see in Table 2 depend largely on the performance of the poststratification estimator.

The major disadvantage of poststratification is that an estimate cannot be produced if any cross-classification cell has no responding elements. Also, the variance cannot be estimated if any cross-classification cell has fewer than two respondents. Furthermore, Särndal et al. (1992) state that, to ensure estimator stability, each group should have a moderate sample size of 20 or more responding elements. If any of the cross-classification groups have smaller sample sizes, then one or more of the component estimates of the poststratification estimator are subject to small sample mean estimation instability, resulting from limited information. Table 3 gives the percentage of samples out of the set of 10,000 for which a poststratified estimate and variance estimate could not be calculated and for which the poststratified estimate should not be considered stable.

Table 3: Percentage of Samples

		$n_S = 500$	$n_S = 2000$	$n_S = 5000$
30%	estimate	99.6%	40.4%	2.2%
	variance	100%	86.5%	13%
	stability	100%	100%	100%
50%	estimate	93.5%	25.4%	1.1%
	variance	100%	65.7%	6.2%
	stability	100%	100%	100%
70%	estimate	67%	3.8%	0.02%
	variance	97.4%	16.7%	0.2%
	stability	100%	100%	100%
80%	estimate	72.6%	6.4%	0.1%
	variance	98.2%	24.9%	0.7%
	stability	100%	100%	100%

From Table 2 we see that generally, purely based on efficiency, the poststratification estimator should be used. However, for surveys with small sample sizes and particularly those with low response rates, Table 3 shows that the special-case calibration estimator is the preferable option since it can be calculated. Additionally, the calibration estimator holds the advantage in cases when the poststratified estimator is not considered stable in that the weight adjustments for individuals in a small cross-classification group do not depend only on the limited information from those selected individuals. Recall that we assume a model in which the response probabilities depend on marginal cross-classification groups membership. So, the calibration method is using pooled marginal information to determine nonresponse adjustments, thus eliminating the small sample estimation instability and resulting in a reliable estimator.

Therefore, the special case calibration estimator is particularly advantageous in the case when one or more of the cross-classification groups are composed of a small portion of the

population. In Table 1 we can see that the group corresponding to State E and Industry 5 contains 0.18% of the overall population. When selecting a simple random sample, the expected sample sizes in that group before nonresponse are 1, 4, and 9 individuals for our three sample sizes, respectively. Therefore, in this type of situation, the poststratified estimator will never be considered reliable. So, while poststratification may result in a more efficient estimator, it will generally not be trusted and, therefore, not used without the collapse of cross-classification cells.

We should note that while poststratification is advantageous with respect to efficiency in many cases, the bias of the special-case calibration estimator is not unreasonable. In fact, over the several cases of response and over the four populations, the bias ranges in absolute value from 119,450 to 1,927,660 while the parameter totals of interest range from 8,363,461 to 9,061,540, all in hundreds of thousands of dollars. Recall that these cases include a sample size of 500 with a response rate of 30%. Therefore, the special-case calibration estimator performs well and should be considered stable without the collapse of cross-classification cells.

5. Further Research

We have evaluated the behavior of the special case of calibration using two-way cross-classification structure characteristics as auxiliary variables as an alternative to poststratification. There are questions that remain that should be treated in future research.

5.1 Incorrect Assumptions

It is possible that we have incorrectly assumed that the response probabilities depend on the marginal cross-classification group information when, in fact, they depend on full cross-classification information. The evaluation of the effects of this incorrect response mechanism assumption on the bias and variance of the proposed estimator would be useful.

Another possibility in practice is to not know the true functional form of the response probabilities and to, therefore, assume incorrectly. It would be useful to examine the effects of using the incorrect functional form in the calibration estimator.

5.2 The Choice of Classification Variables

A major question in practical application of this cross-classified calibration method is: Which of many potential variables should we choose to create the cross-classification structure? Generally, since we assume the same response probability for every element in the same cross-classification group, we would want two variables that separate the individuals in the population into classes that are homogeneous in response tendencies.

Also recall that the variance of the calibration estimator in (31) contains two terms that mostly depend on the cross-classification structure being used. Therefore, a method to choose the auxiliary variables to be used should be developed such that the variance of the calibration estimator is minimized.

5.3 Extending the Cross-Classification Dimension

In the work proposed here, we have limited the cross-classification structure to two classification categories, however, in practical application it is common to have more than two variables of classification that are of interest. For example, in Section 4, we explored the calibration method for the cross-classification of businesses by state and industry type, but adding business size might provide more accurate information.

The way in which the benchmark and model variables are currently defined allow for extension to higher dimensions of classification because of the potential to increase the dimensions of $\mathbf{x}_k$ and $\mathbf{z}_k$ with additional indicator variables. Consider an $A \times B \times C$ triple classification structure. Now, using the same definitions of benchmark and model variables as above, the dimension of the benchmark variables is $P = A \times B \times C$ and the dimension of the model variables is $Q = A + B + C - 2$. Alternatively, different definitions of the model variable can be introduced. As opposed to using indicator variables for the marginal classifications of all three classification categories, the model variables could instead be defined as the marginal classifications of two of the three categories, yielding $Q = A + B - 1$, for example. The cases in which different definitions of the model variables provide the most benefit need to be explored. Introducing different possible response mechanisms depending on the varying definitions of the model variables, however, also leads to the need to evaluate the effect of incorrectly assuming a particular response mechanism.

However, in cases in which the dimensions of $\mathbf{x}_k$ and $\mathbf{z}_k$ are large in comparison to the sample size, the estimation of $\hat{\boldsymbol{\beta}}$ may become unstable or even impossible. In this case or in cases where the marginal double classification structures significantly differ across the third classification categories, a stratified model may provide a better estimate of the parameter of interest.

One advantage of a stratification model is that many assumptions about the relationship between the H strata can be made. For example, the response parameter can be assumed different in each stratum, yielding H independent response parameters, $\boldsymbol{\beta}_h$. Alternatively, the response mechanism can be assumed to be the same for all the strata, yielding one parameter, $\boldsymbol{\beta}$, used in each of the H stratum total estimators. A third possible model is a combination of the two previously described extremes, where subgroups of the strata, denoted $h^* = 1, \ldots, H^*$ with $H^* \leq H$, share the same response mechanism, yielding H^* different response parameters, $\boldsymbol{\beta}_{h^*}$. A few disadvantages of a stratified model are that population marginal double classification structures may need be known for each stratum and that it could prove to be computationally intensive if $\boldsymbol{\beta}_h$ needs to be calculated for all strata.

In the case in which $\hat{\boldsymbol{\beta}}_h$ is estimated separately for each stratum, the overall total estimator is the sum of independent components, meaning that the bias and asymptotic properties of this estimator should reflect the bias and asymptotic properties of its components. In the case in which $\hat{\boldsymbol{\beta}}_{h^*}$ is estimated separately for H^* subgroups of the H strata, the overall estimator is the sum of independent sums of dependent components. Finally, for the case in which a common $\hat{\boldsymbol{\beta}}$ is estimated for all H strata, the components of the overall estimator are completely dependent. For the two cases in which there is dependence between the stratum estimators, the bias and asymptotic properties of the overall estimator will need to be explored. This stratified estimator should also be studied to determine if, in certain cases, there are theoretical efficiency advantages over the estimator using purely expanded classification dimensions.

6. Conclusions

In this paper we have begun the exploration of calibration as a method of nonresponse adjustment in the treatment of two-way cross-classified data. We have determined that the calibration estimator is unbiased and accurate in the case of known response probabilities and asymptotically when the response probabilities must be estimated.

Additionally, from the empirical results we were able to conclude that poststratification is generally a more efficient method to adjust for nonresponse in the case of cross-classified data. However, there were many cases, especially when selecting a small sample and sub-

ject to a low response rate, in which the poststratified estimator could not be calculated without making adjustments that may inhibit interpretation. Additionally, in every case the poststratified estimator would not be considered reliable. Therefore, the calibration method using cross-classification structure characteristics as auxiliary variables is a reasonable alternative that can be considered stable and maintains the interpretive power of the cross-classification variables.

Additional analysis of this special case of the calibration method, as is discussed in Section 5, should be conducted to further evaluate this method. However, given the conclusions resulting from this research, the calibration method is a viable option for nonresponse adjustment for cross-classified data.

REFERENCES

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, Duxbury Press.

Chang, T. (2012). "Calibration Alternatives to Poststratification for Doubly Classified Data," *Survey Methodology*, **38**, 31-41.

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

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). "Generalized Raking Procedures in Survey Sampling," *Journal of the American Statistical Association*, **87**, 376-382.

Deville, J.-C. and Särndal, C.-E. (1992). "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, **88**, 1013-1020.

Estevao, V. and Särndal, C.-E. (2002). "The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling," *Journal of Official Statistics*, **18**, 233-255.

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

Folsom, R. E. (1991). "Exponential and Logistic Weight Adjustment for Sampling and Nonresponse Error Reduction," In *Proceedings of the Social Statistics Section*, 197-202. Washington, DC: American Statistical Association.

Folsom, R. E. and Singh, A. C. (2000). "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification," In *Proceedings of the Survey Research Methods Section*, 598-603. Washington, DC: American Statistical Association.

Fuller, W. A., Loughin, M. M. and Baker, H. D. (1994). "Regression Weighting for the 1987-88 National Food Consumption Survey," *Survey Methodology*, **20**, 75-85.

Kalton, G. and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data," *Survey Methodology*, **12**, 1-16.

Kott, P. S. and Chang, T. (2009). "Using Calibration to Adjust for Nonignorable Unit Nonresponse," *Journal of the American Statistical Association*, **105**, 1265-1275.

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

Kott, P. S. (2005). "Randomization-Assisted Model-Based Survey Sampling," *Journal of Statistical Planning and Inference*, **129**, 263-277.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill Irwin.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*, New York: Springer.

Lundström, S. and Särndal, C.-E. (1999). "Calibration as a Standard Method for the Treatment of Nonresponse," *Journal of Official Statistics*, **15**, 305-397.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

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

Särndal, C.-E. and Lundström, S. (2008). "Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator," *Journal of Official Statistics*, **24**, 167-191.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, New York: Wiley.

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