Using Calibration Weighting to Adjust for Nonresponse or Coverage Error Under a Plausible Model

Ted Chang
University of Virginia
National Agricultural Statistical Service
tcc8v@virginia.edu

Phillip S Kott
National Agricultural Statistical Service
Phil_Kott@nass.usda.gov

Abstract

Calibration forces the weighted estimates of the certain variables to match known or alternatively estimated population totals called benchmarks. It can be used to correct for sample-survey nonresponse or for coverage error resulting from frame undercoverage or unit duplication. The quasi-randomization theory supporting these increasingly important uses treats response or coverage as an additional phase of random sampling, one that takes place before the sample is drawn in the case of coverage errors or after in the case of nonresponse. The functional form of a quasi-random response or coverage model is assumed to be known, while its parameter values are estimated implicitly through the creation of calibration weights. Unfortunately, the variables in a plausible quasi-random model are not necessarily the same as the variables for which benchmark totals are available, as must be commonly be assumed when calibration is used in this manner. Moreover, it may be prudent to keep the number of explanatory variables in a model small. We will address using calibration to adjust for nonreponse or coverage error when the model variables and benchmark variables are allowed to differ as long as the number of benchmark variables is at least as great as the number of model variables. Data from National Agricultural Statistical Service's 2002 Census of Agriculture and simulations based upon that data will be used to illustrate alternative adjustments for nonresponse.

KEY WORDS: Benchmark; Consistency; Coverage model; Back-link function; Quasi-randomization; Response model.

1 Introduction

Calibration weighting ensures that sample-estimated totals of certain calibration or *benchmark* variables match previously determined population totals. Two related special cases of calibration, poststratification and weighting-class adjustment (see Lohr 1999, pp. 266-267) are used extensively to adjust for survey nonresponse, a subject of growing interest as the response rates in both government and private surveys decline.

Oh and Scheuren (1983) provide a theoretical justification for weighting-class adjustment by treating response as a second phase of sample selection. In this *quasi-randomization* (or quasi-design-based) framework, each sampled unit within the same weighting class has an equal and independent probability of selection into the respondent subsample. That probability is estimated implicitly in the weighting process. The prefix "quasi" is added to "randomization" to emphasize that inference depends on an assumed response model.

More complex calibration weighting schemes are proposed by Folsom (1991), Fuller, Louglin, and Baker (1994), and Kott (2004a). In each the probability of response is assumed to be a known *back-link function* of an unknown (but estimatable) linear combination of model variables. The back-link function is the back transformation of the link function in a generalized linear model. See, for example, McCullagh and Nelder (1989). What we have called the "back-link" is sometimes called the "inverse link" in the generalized-linear-model literature.

In Fuller *et al.*, the back-link function has the form $p(\eta) = 1/(1+\eta)$. This allows calibration to have its conventional linear form. Lundström and Särndal (1999) also proposes using calibration in conventional linear form to adjust for nonresponse but without specifying a back-link function.

Folsom proposes more plausible functions for the modeling of response than $p(\eta) = 1/(1+\eta)$. One such is the logistic: $p(\eta) = [1 + exp(-\eta)]^{-1}$. In addition, raking is shown to be a form of calibration weighting with a back-link function of the form $p(\eta) = exp(\eta)$. A follow-up, Folsom and Singh (2002), proposes a class of reasonable back-link functions and points out that calibration can also be used to adjust for coverage errors due to frame undercoverage and/or duplication by assuming a quasi-random coverage model in which the expected number of times a population unit is in the sampling frame is a function of a linear combination of model variables. Although coverage errors takes place before the random sample is drawn and nonresponse after, the same theory applies because the response (coverage) mechanism is assumed to be independent of the sampling mechanism.

In both Fuller *et al.* and Folsom the model variables used to estimate implicitly the probabilities of response are the same as the calibration variables for which one has benchmark totals. In Kott, that is no longer the case. This extension does not require that model-variable totals be known. For example, one can separate respondents into response groups in an analogue to poststratification based on their *survey responses*. Still, Kott assumes the number of model and benchmark variables are equal. Moreover, that paper does not demonstrate the practicality of its approach with data.

Särndal and Lundström (2005) also treats the case where the model and benchmark variables can differ in definition but not in number. In addition, it allows some of the benchmark totals to be calculable from the sample before nonresponse.

The back-link function is not specified. Moreover, the authors do not appear to notice that calibration is possible when model-variable values are known only for the respondents.

We will show how to use calibration to adjust for nonresponse, and by easy inference coverage error, when the number of benchmark variables is at least as great as the number of model variables. In Section 2 we will introduce our notation and motivate our approach to calibration, which is discussed in more detail in Sections 3 and 4.

In section 5 we turn to the estimation of a total for a vector of variables of interest that typically do not include the benchmark variables (since they are either known of previously estimated) but may include some of the model variables. We show how to measure both the additional asymptotic variance due to the nonresponse in a calibration-weighted estimator and the full asymptotic variance of the estimator itself. Section 6 relates our analysis to the existing literature.

Section 7 contains applications of our methodology to nonresponse adjustment for the 2002 Census of Agriculture. We show here how the probability of a farm's responding to the census can be assumed to a function of its survey-reported sales rather than its expected sales before enumeration, as is currently assumed in practice. In section 8 we report on simulations from an artificial population constructed from the respondents of the Census of Agriculture.

Section 9 offers some concluding remarks with a particular emphasis on extending the analysis of the preceding sections to adjusting for coverage errors. This is a topic of increasing interest for the analysis of surveys based on incomplete frames such as telephone and internet surveys.

2 Notation and Motivation

Suppose $\mathbf{z}_i$ is a P-vector of calibration or *benchmark* variables for the i-th population unit, and $\mathbf{x}_i$ a Q-vector of explanatory or *model* variables. We will assume that the probability of i responding (when the unit is selected for the sample) is $p(\mathbf{x}_i'\beta_*)$ for some vector parameter β_*, where $g(\eta) = 1/p(\eta)$ is a known and everywhere monotonic and twice differentiable function, such as $g(\eta) = (1 + exp(-\eta))$.

If β_* were also known, then an expansion estimator for the vector of totals of the benchmark variables would be

$$\widehat{t}_{\mathbf{z}}(\beta_*) = \sum_{i\in\mathcal{R}} \frac{d_i}{p(\mathbf{x}_i'\beta_*)}\mathbf{z}_i, \tag{1}$$

where d_i is the sampling weight in the absence of nonresponse, and $\mathcal{R}$ is the set of respondents.

If $t_{\mathbf{z}}$ is a vector of calibration target values consisting of known, or previously estimated, population totals, then β_* could be estimated from the data using the calibration equation(s)

$$t_{\mathbf{z}} = \sum_{i\in\mathcal{R}} \frac{d_i}{p(\mathbf{x}_i'\widehat{\beta})}\mathbf{z}_i. \tag{2}$$

If the number P of benchmark variables equals the number Q of model variables, equations (2) will usually be sufficient to determine $\widehat{\beta}$. On the other hand, if $P < Q$, $\widehat{\beta}$ will be underdetermined by (2).

Realize, however, that even were β_* known, it is unlikely that $\widehat{t}_{\mathbf{z}}(\beta_*)$ would equal $t_{\mathbf{z}}$ exactly due to sampling variability. The vectors $\widehat{t}_{\mathbf{z}}(\beta_*)$ and $t_{\mathbf{z}}$ should nonetheless be close. With this in mind, we suggest that (1), and its child (2), be viewed as nonlinear regression-type equations

$$t_{\mathbf{z}} = \widehat{t}_{\mathbf{z}}(\beta) + \epsilon \tag{3}$$

where

$$\widehat{t}_{\mathbf{z}}(\beta) = \sum_{i \in \mathcal{R}} \frac{d_i}{p(\mathbf{x}_i'\beta)} \mathbf{z}_i$$

and ϵ is a P-vector of "errors". In the nonlinear regression paradigm, it is desirable that $P > Q$ and, indeed, the more calibration targets the merrier. The parameter β can be estimated by minimizing an objective function of the form

$$\rho(\beta) = (t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta))' \mathbf{W} (t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta)) \tag{4}$$

for some appropriately chosen $P \times P$ positive definite matrix $\mathbf{W}$. We show in the next section that minimizing the objective function will yield consistent estimators for the parameters no matter what one chooses for $\mathbf{W}$ under mild conditions. Nevertheless, it stands to reason that some choices will lead to more efficient estimators than others.

The nonlinear regression formulation of (3) suggests setting $\mathbf{W} = \mathbf{V}^{-1}$ for some suitably defined variance matrix $\mathbf{V}$ of ϵ. When this choice of $\mathbf{W}$ depends on β, we propose an iterative procedure analogous to what would be used with a fixed $\mathbf{W}$. Given a guess $\widehat{\beta}_0$ of β_*, we can linearize the regression (3) at $\widehat{\beta}_0$. The solution to the linearized regression is the next guess $\widehat{\beta}_1$. This procedure is described more thoroughly in Section 3.

An obvious candidate for $\mathbf{V}$ is $\widehat{Var}_{qr}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*)$, an estimator for the quasi-randomization variance of $\widehat{t}_{\mathbf{z}}(\beta)$ assuming $\beta = \beta_*$. If the sampling scheme is without replacement so that $\widehat{t}_{\mathbf{z}}(\beta)$ of (3) is the Horvitz Thompson estimator, then one such quasi-randomization variance estimator is

$$\widehat{Var}_{qr}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*) = \sum_{i,j \in \mathcal{R}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j p_i p_j} \mathbf{z}_i \mathbf{z}_j' + \sum_{i \in \mathcal{R}} \frac{1 - p_i}{p_i^2 \pi_i} \mathbf{z}_i \mathbf{z}_i', \tag{5}$$

where $p_i = p(\mathbf{x}_i'\beta)$.

In the quasi-randomization framework supporting (5) one assumes the set $\mathcal{R}$ of respondents results from a two-phase sample of the target population $\mathcal{U}$. In the first phase, the sample $\mathcal{S}$ is drawn without replacement from the population with inclusion probabilities $\pi_i = Pr[\ i \in \mathcal{S}]$ and $\pi_{ij} = Pr[\ i, j \in \mathcal{S}]$. In this case, $d_i = \pi_i^{-1}$. In the second phase, $\mathcal{R}$ is a Poisson subsample of $\mathcal{S}$ with unit selection probabilities of the form $p_i = p(\mathbf{x}_i'\beta)$.

If the targets $t_{\mathbf{z}}$ are previously estimated population totals, then it is be reasonable to compute

$$\mathbf{V} = \widehat{Var}_{qr}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*) + \widehat{Var}(t_{\mathbf{z}})$$

where $\widehat{Var}(t_{\mathbf{z}})$ is a good estimate of the variance of $t_{\mathbf{z}}$. In some applications $\widehat{Var}(t_{\mathbf{z}})$ may be much greater than $\widehat{Var}_{qr}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*)$, and it will be reasonable to set $\mathbf{W}$ to $(\widehat{Var}(t_{\mathbf{z}}))^{-1}$, which is not a function of β at all.

A common procedure correcting for nonresponse is poststratification. As we noted previously, poststratifcation is a special case of calibration weighting. Let $P = Q$ be the number poststrata and let $\mathbf{z} = \mathbf{x}$ be P-vectors of indicator variables for the P poststrata. In this case $t_{\mathbf{z}} = [N_1 \cdots N_P]^{'}$, where N_h is the population size of stratum h. Equation (2) has the exact solution

$$\begin{aligned} p(\widehat{\beta}_h) &= (\sum_{i \in \mathcal{R}_h} d_i)/N_h \\ a_i &= 1/p(\widehat{\beta}_h) \text{ when } i \in \mathcal{R}_h \\ \widehat{\beta} &= [\widehat{\beta}_1 \cdots \widehat{\beta}_P]^{'}, \end{aligned}$$

where $\mathcal{R}_h$ is the set of respondents in stratum h, a_i is the (calibration) adjustment factor for unit i, and $c_i = a_i d_i$ is the unit's calibration weight. If the sampling design is a simple random sample, then $d_i = N/n$ for all sampled units, and the calibration weight is $c_i = a_i d_i = N_h/n_h$ for $i \in \mathcal{R}_h$.

We are assuming when using poststratication to adjust for nonresponse that all units in the h-th poststratum or *response group* have an equal probability of responding. In many situations, however, a more reasonable response model would employ a vector $\mathbf{x}$ of variables different from the vector $\mathbf{z}$ of benchmark variables when forming these groups. Moreover, the variables in $\mathbf{x}$ need not have known or alternatively estimated population totals. This is the case in the example discussed in Section 7.

3 Some Details

Given a guess $\widehat{\beta}_0$ of β and matrix $\mathbf{W}(\widehat{\beta}_0)$ we linearize (3) at $\widehat{\beta}_0$ and obtain

$$t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta}_0) \approx \widehat{\mathbf{H}}(\widehat{\beta}_0) \left(\beta - \widehat{\beta}_0\right) + \epsilon, \tag{6}$$

where $\widehat{\mathbf{H}}(\widehat{\beta}_0)$ is the $P \times Q$ matrix

$$\widehat{\mathbf{H}}(\widehat{\beta}_0) = \frac{\partial \widehat{t}_{\mathbf{z}}(\widehat{\beta}_0)}{\partial \beta} = \sum_{i \in \mathcal{R}} d_i g_1(\mathbf{x}_i^{'}\widehat{\beta}_0)\mathbf{z}_i\mathbf{x}_i^{'}, \tag{7}$$

and $g_1(\mathbf{x}_i^{'}\widehat{\beta}_0)$ is the first derivative of $g(\eta) = 1/p(\eta)$ evaluated at $\mathbf{x}_i^{'}\widehat{\beta}_0$.

The (weighted) linear regression estimate $\widehat{\beta}_1$ corresponding to (6) minimizes the objective function $\mathbf{U}^{'}\mathbf{W}(\widehat{\beta}_0)\mathbf{U}$ where $\mathbf{U} = t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta}_0) - \widehat{\mathbf{H}}(\widehat{\beta}_0)(\beta - \widehat{\beta}_0)$. It is given by the *update equation*:

$$\widehat{\beta}_1 = \widehat{\beta}_0 + \left[\widehat{\mathbf{H}}(\widehat{\beta}_0)^{'}\mathbf{W}(\widehat{\beta}_0)\widehat{\mathbf{H}}(\widehat{\beta}_0)\right]^{-1}\left[\widehat{\mathbf{H}}(\widehat{\beta}_0)^{'}\mathbf{W}(\widehat{\beta}_0)\left(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta}_0)\right)\right]. \tag{8}$$

For simplicity, we assume $\widehat{\mathbf{H}}(\beta)$ and $\mathbf{W}(\beta)$ are of full rank everywhere. This will allow us to always be able to invert matrices when the need arises.

Iteration continues with $\widehat{\beta}_1$ serving the role of $\widehat{\beta}_0$ in (8), and so on until we reach a step K, if such a step can be reached, where

$$\widehat{\mathbf{H}}^{'}\mathbf{W}\left(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta})\right) = 0 \tag{9}$$

with the matrices $\mathbf{H}$ and $\mathbf{W}$ evaluated at $\widehat{\beta} = \widehat{\beta}_K$. As the iteration starting with (8) is a Newton-Raphson type method, it is sometimes helpful to limit the size of the step from, say, $\widehat{\beta}_{k-1}$ to $\widehat{\beta}_k$ for some $k = 1, ..., K$. What this means is that when $\left[\widehat{\mathbf{H}}(\widehat{\beta}_{k-1})\mathbf{W}(\widehat{\beta}_{k-1})\widehat{\mathbf{H}}(\widehat{\beta}_{k-1})\right]^{-1}\left[\widehat{\mathbf{H}}(\widehat{\beta}_{k-1})'\mathbf{W}(\widehat{\beta}_{k-1})\left(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta}_{k-1})\right)\right]$ is deemed too large, it can be replaced with some fraction of itself in the update equation.

If $\mathbf{W}$ is the inverse of an estimate for the variance matrix of $\widehat{t}_{\mathbf{z}}(\widehat{\beta})$, an alternative derivation and justification of the update equation (8) follows from Thompson (1997). Consider the equations $\widehat{t}_{\mathbf{z}}(\beta) - t_{\mathbf{z}} = 0$ as P estimating equations for the Q coefficients β. If $\mathbf{A}$ is a $Q \times P$ matrix of constants, let $\widehat{\beta}_A$ denote the solution to the estimating equations

$$\mathbf{A}\widehat{t}_{\mathbf{z}}(\beta) = \mathbf{A}t_{\mathbf{z}}. \tag{10}$$

A choice for $\mathbf{A}$ such that $\widehat{\beta}_A$ has a minimum asymptotic variance is $\mathbf{A}^* = \widehat{\mathbf{H}}(\widehat{\beta}_{A^*})'[Var(\widehat{t}_{\mathbf{z}}(\widehat{\beta}_{A^*}))]^{-1}$, where $\widehat{\beta}_{A^*}$ results from the convergence of update equation (8).

Observe that when $\mathbf{W} = \mathbf{W}(\beta)$ depends upon β, our suggested procedure for estimating β_* does *not* minimize the objective function (4). This is because if (4) were differentiated with respect to β, then there would be a term for the derivative of $\mathbf{W}$ which is not accounted for in the linearization (6). In other words, letting

$$\rho_0(\beta, \gamma) = (t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta))'\mathbf{W}(\gamma)(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta)),$$

full minimization would solve the equation

$$0 = \frac{\partial \rho_0}{\partial \beta}(\widehat{\beta}, \widehat{\beta}) + \frac{\partial \rho_0}{\partial \gamma}(\widehat{\beta}, \widehat{\beta}) \tag{11}$$

whereas the partial minimization obtained through iterated use of (8) sets only the first term of the right hand side of (11) to zero.

We have adopted a partial minimization approach here because employing an estimator for β_* that solves (9) plays a crucial role in measuring the added variance from nonresponse in a calibration-estimated total as we will see in Section 5. For now, we show that our choice for $\widehat{\beta}$ is consistent given the quasi-random model under mild conditions.

Since $g(\eta) = 1/p(\eta)$ is everywhere monotonic and twice differentiable, there is a θ_i between $\mathbf{x}_i'\widehat{\beta}$ and $\mathbf{x}_i'\beta_*$ such that $g(\mathbf{x}_i'\widehat{\beta}) - g(\mathbf{x}_i'\beta_*) = g_1(\theta_i)\mathbf{x}_i'(\widehat{\beta} - \beta_*)$. Thus, we can derive the following equality from (9):

$$\widehat{\beta} = \beta_* + \left[\widehat{\mathbf{H}}'\mathbf{W}\widehat{\mathbf{H}}_\theta\right]^{-1}\left[\widehat{\mathbf{H}}'\mathbf{W}\left(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta_*)\right)\right], \tag{12}$$

where

$$\widehat{\mathbf{H}}_\theta = \sum_{i \in \mathcal{R}} d_i g_1(\theta_i)\mathbf{z}_i\mathbf{x}_i'.$$

For simplicity, we assume $\widehat{\mathbf{H}}_\theta$, like $\widehat{\mathbf{H}} = \widehat{\mathbf{H}}(\widehat{\beta})$ and $\mathbf{W} = \mathbf{W}(\widehat{\beta})$, has full rank.

An asymptotic structure is required to analyze the large-sample behavior of $\widehat{\beta}$. Formally, one must postulate a sequence of populations and sampling designs so that the population size N can grow along with the sample size n. For some

designs, the realized sample size is random and hence n must be interpreted as an expected sample size.

From (12), it is not hard to see that if

1. $\widehat{\mathbf{H}}/N$, $\widehat{\mathbf{H}}_\theta/N$, and $\mathbf{W}$ or a multiple of $\mathbf{W}$ - such as $(N^2/n)\mathbf{W}$ - converge in probability, at the rate $n^{-1/2}$, to matrices of full rank, and if
2. $\sqrt{(n)}(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\beta_*))/N$ converges in probability,

then $\widehat{\beta}$ converges in probability to β_*. Furthermore

$$g(\mathbf{x}_i'\widehat{\beta}) - g(\mathbf{x}_i'\beta_*) = g_1(\mathbf{x}_i'\widehat{\beta})\mathbf{x}_i'(\widehat{\beta} - \beta_*) + \frac{1}{2}g_2(\xi_i)[\mathbf{x}_i'(\widehat{\beta} - \beta_*)]^2, \tag{13}$$

for some ξ_i between $\mathbf{x}_i'\widehat{\beta}$ and $\mathbf{x}_i'\beta_*$, where $g_2(.)$ is the second derivative of g. This equality will prove helpful in Section 5.

To see the reasonableness of these conditions, it is helpful to remember that $\widehat{\mathbf{H}}$ and $\widehat{\mathbf{H}}_\theta$ are estimates of population totals. Proofs of these assumptions require rather technical conditions on the sequence of populations and sampling designs.

An example of suitable conditions can be found in Fuller and Isaki (1981). This paper can be applied to many useful without replacement designs. In their notation the inclusion probabilities for the r-th universe and design are denoted by $\pi_{i(r)}$ and $\pi_{ij(r)}$. In our context, these are probabilities for inclusion in the respondent sample $\mathcal{R}$. Retaining our notation of π_i and π_{ij} as the inclusion probabilities for the full sample $\mathcal{S}$, and ignoring in this notation the dependence of these probabilities on r, the Fuller-Isaki inclusion probabilities are related to those of this paper by $\pi_{i(r)} = \pi_i p(\mathbf{x}_i'\beta_*)$ and $\pi_{ij(r)} = \pi_{ij}p(\mathbf{x}_i'\beta_*)p(\mathbf{x}_j'\beta_*)$.

4 Some Choices for W

When $P = Q$, equation (8) reduces to

$$\widehat{\beta}_1 = \widehat{\beta}_0 + \widehat{\mathbf{H}}^{-1}\left(t_{\mathbf{z}} - \widehat{t}_{\mathbf{z}}(\widehat{\beta}_0)\right). \tag{14}$$

Thus, in this case, the form of $\mathbf{W}$ is irrelevant for the update equation. Indeed (14) is the Newton-Raphson update equation for solving the equation $t_{\mathbf{z}} = \widehat{t}_{\mathbf{z}}(\widehat{\beta})$, and the solution, if it exists, will also minimize (4).

When $P > Q$, we proposed setting $\mathbf{W}(\beta)$ to the inverse of $\mathbf{V} = \widehat{Var}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*) + \widehat{Var}(t_{\mathbf{z}})$, where the latter term is provided to us from external sources (and may be $\mathbf{0}$). If the sample $\mathcal{S}$ is selected without replacement, then (5) is an unbiased estimate of $Var(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*)$. This is an immediate consequence of equation (9.3.7) in Särndal *et al.*(1992).

For a stratified multistage design with primary sampling units (PSU's) chosen *with* replacement, let $h = 1, \ldots, H$ index the strata. Suppose n_h PSU's are chosen with replacement in stratum h. Let $n_h q_{hi}$ be the expected number of times PSU i from stratum h will appear in the first stage sample

Changing the meaning of i slightly, let $\mathcal{R}_{hi}$ be the set of subsampled and then responding elements in the i-th PSU selected from stratum h, and let d_{hij} be the

sampling weight (before nonresponse adjustment) for element $j \in \mathcal{R}_{hi}$. Define

$$\widehat{t}_{\mathbf{z}hi}(\beta) = n_h q_{hi} \sum_{j \in \mathcal{R}_{hi}} \frac{d_{hij}}{p(\mathbf{x}'_{hij}\beta)} \mathbf{z}_{hij}, \text{ and}$$

$$\widehat{t}_{\mathbf{z}h}(\beta) = n_h^{-1} \sum_{i=1}^{n_h} q_{hi}^{-1} \widehat{t}_{\mathbf{z}hi}(\beta).$$

Then $\widehat{t}_{\mathbf{z}}(\beta) = \sum_h \widehat{t}_{\mathbf{z}h}(\beta)$ is an unbiased estimate of $t_{\mathbf{z}}$. When $\beta = \beta_*$ its quasi-randomization variance can be unbiasedly estimated by

$$\widehat{Var}_{qr}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*) = [n_h(n_h-1)]^{-1} \sum_{i=1}^{n_h} (q_{hi}^{-1}\widehat{t}_{\mathbf{z}hi}(\beta) - \widehat{t}_{\mathbf{z}h}(\beta))(q_{hi}^{-1}\widehat{t}_{\mathbf{z}hi}(\beta) - \widehat{t}_{\mathbf{z}h}(\beta))'.$$

This follows from assuming response is additional phase of subsampling, one that is independent across all sampled elements.

Finally, notice that when $\widehat{Var}(t_{\mathbf{z}}) = \mathbf{0}$, the estimate $\widehat{\beta}$ is unchanged when $Var_{qr}(\widehat{t}_{\mathbf{z}}(\beta))$ is only estimated up to a multiplicative constant within $\mathbf{W}^{-1}$. This suggests invoking the spirit of design effects (Kish 1965) and using the estimated quasi-randomization variance of what would have resulted from a with-replacement simple random sample of size n drawn from a population of size N:

$$\widehat{Var}_{srs}(\widehat{t}_{\mathbf{z}}(\beta)|\beta = \beta_*) = \frac{N^2}{n(N-1)} \sum_{i \in \mathcal{R}} d_i p_i^{-1} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})',$$

where

$$\overline{\mathbf{z}} = (\sum_{i \in \mathcal{R}} d_i p_i^{-1})^{-1} \sum_{i \in \mathcal{R}} d_i p_i^{-1} \mathbf{z}_i,$$

and, as before, $p_i = p(\mathbf{x}'_i\beta)$. In practice, the scalar multiple $\frac{N^2}{n(N-1)}$ can be dropped from $\widehat{\mathbf{V}}$ within $\mathbf{W} = \widehat{\mathbf{V}}^{-1}$.

5 The Calibration-Weighted Estimator for a Population Total and Its Quasi-randomization Variance

Let $\mathbf{y}_i$ be a vector of variables of interest. The purpose of calibration weighting is to estimate population totals for the components of $\mathbf{y}_i$ or functions of those totals, like population ratios. Typically, the components of $\mathbf{y}_i$ will not include the benchmark variables in $\mathbf{z}_i$ as they are already known or have been alternatively estimated. They may include model variables in $\mathbf{x}_i$. In principle, they could include both types of variables.

Our calibration-weighted estimator for the total $t_{\mathbf{y}}$ is

$$\widehat{t}_{\mathbf{y}}(\widehat{\beta}) = \sum_{i \in \mathcal{R}} \frac{d_i}{p(\mathbf{x}'_i\widehat{\beta})} \mathbf{y}_i = \sum_{i \in \mathcal{R}} d_i g(\mathbf{x}'_i\widehat{\beta}) \mathbf{y}_i = \sum_{i \in \mathcal{R}} c_i \mathbf{y}_i, \tag{15}$$

where $c_i = d_i g(\mathbf{x}'_i\widehat{\beta})$ is the calibration weight for unit i. Define

$$\widehat{\mathbf{H}}_{\mathbf{y}} = \frac{\partial \widehat{t}_{\mathbf{y}}}{\partial \beta}(\widehat{\beta}) = \sum_{i \in \mathcal{R}} d_i g_1(\mathbf{x}'_i\widehat{\beta}) \mathbf{y}_i \mathbf{x}'_i, \text{ and}$$

$$\mathbf{B} = \widehat{\mathbf{H}}_{\mathbf{y}} \left[\widehat{\mathbf{H}}' \mathbf{W} \widehat{\mathbf{H}}\right]^{-1} \widehat{\mathbf{H}}' \mathbf{W},$$

where $\widehat{\mathbf{H}} = \widehat{\mathbf{H}}_{\mathbf{z}}$ is the matrix of partial derivatives (7) which, like $\mathbf{W}$, is evaluated at $\widehat{\beta}$. We assume that the components of $\widehat{\mathbf{H}}_{\mathbf{y}}/N$ are $O_P(1)$.

Writing $\mathbf{y}_i$ as $\mathbf{B}\mathbf{z}_i + (\mathbf{y}_i - \mathbf{B}\mathbf{z}_i)$, we have

$$\begin{aligned}\widehat{t}_{\mathbf{y}}(\widehat{\beta}) &= \mathbf{B}\widehat{t}_{\mathbf{z}}(\widehat{\beta}) + \sum_{i\in\mathcal{R}} \frac{d_i}{p(\mathbf{x}_i'\widehat{\beta})}(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i) \\ &= \mathbf{B}t_{\mathbf{z}} + \sum_{i\in\mathcal{R}} d_i g(\mathbf{x}_i'\widehat{\beta})(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i).\end{aligned}$$

The latter equality follows from (9). Continuing by applying the equality in (13), we have

$$\begin{aligned}\widehat{t}_{\mathbf{y}}(\widehat{\beta}) = \mathbf{B}t_{\mathbf{z}} &+ \sum_{i\in\mathcal{R}} d_i g(\mathbf{x}_i'\beta_*)(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i) \\ &+ (\widehat{\mathbf{H}}_{\mathbf{y}} - \mathbf{B}\widehat{\mathbf{H}})(\widehat{\beta} - \beta_*) + \frac{1}{2}\sum_{i\in\mathcal{R}} d_i g_2(\xi_i)[\mathbf{x}_i'(\widehat{\beta} - \beta_*)]^2(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i) \\ = \mathbf{B}t_{\mathbf{z}} &+ \sum_{i\in\mathcal{R}} d_i g(\mathbf{x}_i'\beta_*)(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i) + \mathbf{O}_P(N/n),\end{aligned}$$

assuming all the components of $\sum_{i\in\mathcal{R}} d_i|g_2(\xi_i)|(\mathbf{x}_i'\mathbf{x}_i)|\mathbf{y}_i - \mathbf{B}\mathbf{z}_i|$ are $O_P(N)$ (where $|\mathbf{q}|$ is a vector containing the absolute values of the components of $\mathbf{q}$). This leads to

$$(\widehat{t}_{\mathbf{y}}(\widehat{\beta}) - t_{\mathbf{y}})/N = (\widehat{t}_{\mathbf{u}}(\beta_*) - t_{\mathbf{u}})/N + \mathbf{O}_P(1/n), \tag{16}$$

where $\mathbf{u}_i = \mathbf{y}_i - \mathbf{B}\mathbf{z}_i$.

Unfortunately, u_i depends on values for responding units other than i through $\mathbf{B}$, so $t_{\mathbf{u}}$ is not really a vector of population totals. By adding the mild assumption that as n grows arbitrarily large, $\mathbf{B} - \mathbf{B}_* = \mathbf{O}_P(1/n^{1/2})$, where $\mathbf{B}_*$ is the probability limit of $\mathbf{B}$, we can infer that $(\widehat{t}_{\mathbf{z}}(\beta_*) - t_{\mathbf{z}})(\mathbf{B} - \mathbf{B}_*)/N = \mathbf{O}_P(1/n)$. Thus, we can replace the $\mathbf{u}_i$ in (16) by $\mathbf{u}_{i*} = \mathbf{y}_i - \mathbf{B}_*\mathbf{z}_i$ and still have

$$(\widehat{t}_{\mathbf{y}}(\widehat{\beta}) - t_{\mathbf{y}})/N = (\widehat{t}_{\mathbf{u}*}(\beta_*) - t_{\mathbf{u}*})/N + \mathbf{O}_P(1/n). \tag{17}$$

That is to say, $Var_{qr}(N^{-1}\widehat{t}_{\mathbf{y}}(\widehat{\beta})) = Var_{qr}(N^{-1}\widehat{t}_{\mathbf{u}*}(\beta_*)) + O(n^{-3/2})$.

Recall that $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ has made use of a quasi-random model to adjust for nonresponse. The variance of the respondent indicator function for sampled unit i (which is 1 if i responds and 0 otherwise) is $p(\mathbf{x}_i'\beta_*)(1 - p(\mathbf{x}_i'\beta_*))$. It is not difficult to see from (17) that under the quasi-random model and other mild assumptions we have made, the component of the randomization variance of $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ due to nonresponse can be estimated in an asymptotically unbiased manner by

$$\widehat{Var}_{add} = \sum_{i\in\mathcal{R}} \frac{d_i^2}{p(\mathbf{x}_i'\widehat{\beta})^2}(1 - p(\mathbf{x}_i'\widehat{\beta}))(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i)(\mathbf{y}_i - \mathbf{B}\mathbf{z}_i)'. \tag{18}$$

when $t_{\mathbf{z}}$ is a known vector of population values. (Note that there is no asymptotic effect from our using $\widehat{\beta}$ and $\mathbf{B}$ in place of β_* and $\mathbf{B}_*$ in (18).)

The right hand side of (18) does not have a component that accounts for the uncertainty in estimating β_*. This is because the way β_* is estimated (i.e., through (9)) and $\mathbf{B}$ is calculated leads to (16), where the additional error caused estimating β_* by $\widehat{\beta}$ is asymptotically ignorable.

When the original sample is drawn randomly without replacement, and $t_{\mathbf{z}}$ is a known vector of population totals, a reasonable estimator for the quasi-randomization variance of $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ can be computed using the right hand side of (5) with p_i set to $p(\mathbf{x}_i^{'}\widehat{\beta})$ and $\mathbf{z}_i$ replaced by $\mathbf{u}_i$. That is to say,

$$\widehat{Var}_{qr}(\widehat{t}_{\mathbf{y}}(\widehat{\beta})) = \sum_{i,j\in\mathcal{R}} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j p_i p_j}\mathbf{u}_i\mathbf{u}_j^{'} + \sum_{i\in\mathcal{R}} \frac{1-p_i}{p_i^2\pi_i}\mathbf{u}_i\mathbf{u}_i^{'}, \tag{19}$$

where $p_i = p(\mathbf{x}_i^{'}\widehat{\beta})$. Proving $\widehat{Var}_{qr}(\widehat{t}_{\mathbf{y}}(\widehat{\beta}))$ is asymptotically unbiased is not trivial when the right hand side of of (19) has $O(n^2)$ terms. See Kim *et al.* for sufficient additional restrictions on the coefficients in that situation.

Similarly, when $t_{\mathbf{z}}$ is a known vector of population values and the original sample is drawn using a stratified multi-stage routine and the first stage drawn *with* replacement one can compute the variance estimator for $\widehat{t}_{\mathbf{y}}$ under a first-stage-with-replacement design as described in Section 4, again with p_i set to $p(\mathbf{x}_i^{'}\widehat{\beta})$ and $\mathbf{u}_i$ replacing $\mathbf{z}_i$.

There is an additional component in the asymptotic variance of $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ when $t_{\mathbf{z}}$ is *not* known with certainty but is estimated from independent external sources. An obvious measure for this component is

$$\widehat{Var}_{ext} = \mathbf{B}\widehat{Var}(t_{\mathbf{z}})\mathbf{B}^{'}, \tag{20}$$

where $\widehat{Var}(t_{\mathbf{z}})$ is an externally-provided estimate for the variance of $t_{\mathbf{z}}$.

6 Discussion

It is not unreasonable to conclude from (18) and (19) that $\mathbf{u}_i$ plays the role of a regression residual in variance estimation. Following that train of thought, $\mathbf{B}$ can be viewed as the generalization of a sample-weighted linear regression coefficient.

This is exactly what $\mathbf{b} = \mathbf{B}^{'}$ is in the special case explored by Fuller *et al.* (1994) where y_i is a scalar, $\mathbf{x}_i = \mathbf{z}_i$, and $g(\eta) = 1 + \eta$. In that paper, the weights in the weighted linear regression coefficient are the sampling weights before nonresponse adjustment (that is, the d_i). Generalizing to a possibly nonlinear $g(\eta)$ merely affects those weights as can be seen in Folsom and Singh (2000), where

$$\mathbf{b} = (\sum_{i\in\mathcal{R}} d_i g_1(\mathbf{x}_i^{'}\widehat{\beta})\mathbf{x}_i\mathbf{x}_i^{'})^{-1}\sum_{i\in\mathcal{R}} d_i g_1(\mathbf{x}_i^{'}\widehat{\beta})\mathbf{x}_i y_i. \tag{21}$$

Observe that $g_1(\eta)$ in (21) is unity in Fuller *et al.*. Moreover, when the components of $\mathbf{x}_i$ are indicators of mutually-exclusive group membership (as in poststratification), the $g_1(\mathbf{x}_i^{'}\widehat{\beta})$ terms can be shown to effectively drop out of the equation.

Lundström and Särndal (1999) computes $\mathbf{b}$ with the function $g(.)$ replacing $g_1(.)$ above. Unfortunately, with this $\mathbf{b}$ the derivation of (16) fails for most back-link functions. To be fair, in Särndal and Lundström (2005) the authors only claim to provide a "rationale" for their variance-estimation approach not a proof (p. 139). Moreover, they argue that getting the form of the back-link function right is rarely important. See the discussion of raking on p. 318 of Lundström and Särndal or p. 75 of Särndal and Lundström.

By severing the connection between the benchmark and model variables while keeping P=Q, the $\mathbf{b}$ in Kott (2004a) becomes

$$\mathbf{b} = (\sum_{i \in \mathcal{R}} d_i g_1(\mathbf{x}_i^{'}\widehat{\beta})\mathbf{x}_i\mathbf{z}_i^{'})^{-1} \sum_{i \in \mathcal{R}} d_i g_1(\mathbf{x}_i^{'}\widehat{\beta})\mathbf{x}_i y_i, \tag{22}$$

a weighted instrumental-variable regression coefficient. In the implicit *prediction model* of the linear regression, y_i remains the variable of interest, the components of $\mathbf{z}_i$ are the explanatory variables, and the components of $\mathbf{x}_i$ serve as the instruments. This role reversal for the benchmark and model variables can plainly be seen in $\mathbf{u}_i = u_i = y_i - \mathbf{b}^{'}\mathbf{z}_i$. Remember, however, that justification for calibration given here rests entirely on a response model governing which units respond rather than a prediction model linking the behavior of any component of $\mathbf{y}_i$ to either $\mathbf{z}_i$ or $\mathbf{x}_i$.

Kott also develops a jackknife for use with a stratified multi-stage sampling routine that avoids iteration in the creation of replicate weights. It is not hard to show using the reasoning in the text that this jackknife retains its asymptotic properties in the $P > Q$ case when the Q-vector $\widetilde{\mathbf{z}}_i = \widehat{\mathbf{H}}^{'}\mathbf{W}\mathbf{z_i}$ replaces the original vector of benchmark variables. Accordingly, $\widehat{t}_{\widetilde{\mathbf{z}}}(\widehat{\beta}) = t_{\widetilde{\mathbf{z}}}$, while $u_i = y_i - \widetilde{\mathbf{b}}^{'}\widetilde{\mathbf{z}}_i$, and $\widetilde{\mathbf{b}}$ is computed using (22) with $\widetilde{\mathbf{z}}_i$ replacing $\mathbf{z}_i$.

The availability of relative simple quasi-randomization variance estimators is one advantage of the calibration approach over an alternative weight-adjustment-for-nonresponse scheme in which a response propensity model is estimated directly from the full sample (respondents and nonrespondents) and then the estimated unit response probabilities treated as Poisson probabilities of response in an expansion estimator aggregated only over the respondents. A more important advantage of the calibration approach developed here is that the response model can include variables with values unknown for the nonrespondents, as we will demonstrated in the next section.

7 Example: the 2002 Census of Agriculture

Despite its name, it is helpful to think of the 2002 US Census of Agriculture as a survey. Its core was indeed a census of all places *with the potential* of producing $1,000 or more of annual agricultural sales - what the National Agricultural Statistics Service (NASS) defines as farms. This core, however, had to be supplemented to two directions. First, not all entities on the Census Mailing List (CML) maintained by NASS responded to the Census of Agriculture. Second, not all places NASS defined as farms were on the CML.

To compensate for nonresponse and undercoverage, reweighting of Census records was used. Unlike most surveys, the completed records in the 2002 Census of Agriculture had original weights of 1. These weights were adjusted first to account for the nonresponse and then the undercoverage. How both of these were done is explained in some detail on the NASS web site (Kott 2004b).

Our attention in this section will be focused on the adjustment for nonresponse in the 2002 Census of Agriculture. NASS used a weighting-class/poststratification approach to adjusting for nonresponse. The agency divided the entities receiving Census forms into mutually exclusive response groups (weighting classes, poststrata) based on what NASS believed to be each entity's county of operation, its size class as

measured by expected sales, and whether or not the entity responded to an agency survey since 1997. Potential farms above a certain expected size or without expected sales information were removed from this categorization and handled separately, as were entities whose forms were returned as undeliverable by the Post Office.

Under the quasi-randomization model implicit when reweighting by response group, every unit in a group is assumed to be equally likely to respond, irrespective of its Census item values. This assumption has one obvious defect when used in the Census of Agriculture. Groups are defined using expected 2002 sales as assigned by NASS before enumeration rather than the actual sales reported on the US Census of Agriculture. It is more reasonable to assume similar behavior from farms in the same actual-sales group than from entities in the sample expected-sales group. An obvious example: two farms in the sample expected-sales group but in different actual-sales groups can have vastly different actual sales, a Census item.

The single most important Census item is whether an entity meets the definition of a farm (the item value is 1 if it does, 0 otherwise). Here, again, actual sales is a much better predictor for the item value than expected sales.

Why then did NASS use expected sales in creating response groups? The answer applies to all surveys currently using a weighting-class approach to nonresponse adjustment not just the Census of Agriculture. In order to apply reweighting the entire sample, or in this case the entire categorizable population, must be assigned to response groups. Since NASS only knew the actual sales for respondents, it could not use actual sales in creating groups.

NASS actually created response groups within counties. When too few potential farms responded in a putative group, collapsing rules were followed. In order to focus on the repercussions of using expected-sales in response-group formation and to avoid sticky small-group problems, the analysis here forms groups at the state level.

Paralleling the within-county routines actually used by NASS, reweighting is first done within each state by creating five mutually exclusive response groups:

- Z-Group 1: Expected 2002 sales less than \$2,500;
- Z-Group 2: Expected 2002 sales between \$2,500 and \$9,999;
- Z-Group 3: Expected 2002 sales between \$10,000 and \$49,999 and previously reported survey data from 1997 or later;
- Z-Group 4: Expected 2002 sales greater than or equal to \$50,000 and reported survey data from 1997 or later;
- Z-Group 5: Expected 2002 sales greater than or equal to \$10,000 and no reported survey data from 1997 or later.

All the units in each state data set belonged to one of these five groups. Mathematically, $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots, z_{i5})'$, where $z_{ih} = 1$ when i is in Z-Group h and $z_{ih} = 0$ otherwise.

Because all the $d_i = 1$, the nonresponse-adjustment (poststratification) weight for unit i is $c_i = a_i = N_h/n_h$ when the unit is in Z-Group h, where N_h is the number of entities on the CML in Z-group h and n_h is the number of responding entities in Z-group h.

Table 1: Fitted weights for the five **x**-variable calibration

	$\widetilde{a}_1$	$\widetilde{a}_2$	$\widetilde{a}_3$	$\widetilde{a}_4$	$\widetilde{a}_5$
CA	1.095	1.118	1.164	1.227	1.352
DE	1.153	1.136	1.531	1.151	1.431
IL	0.999	1.286	1.275	1.239	2.504
LA	1.164	0.954	1.281	1.226	1.670
SD	1.076	0.942	1.143	1.228	1.720

We contrast reweighting with calibration using these five Z-group indicator variables as benchmark variables. We will present here the results from five states. California (CA) and Delaware (DE) have been chosen because they represent diametrically opposite states from an agricultural standpoint. Illinois (IL), Louisiana (LA), and South Dakota (SD) are included because they pose interesting difficulties in the analysis as we shall see. Table 2 displays the benchmark targets $t'_{\mathbf{z}} = (N_1, \cdots, N_5)$.

In one set of calibration runs, we use as model variables the indicator variables for five groups defined using *actual* sales, as reported on the Census of Agriculture. These five groups are:

- X-Group 1: Actual 2002 sales less than \$2,500;
- X-Group 2: Actual 2002 sales between \$2,500 and \$9,999;
- X-Group 3: Actual 2002 sales between \$10,000 and \$49,999 and previously reported survey data from 1997 or later;
- X-Group 4: Actual 2002 sales greater than or equal to \$50,000 and reported survey data from 1997 or later;
- X-Group 5: Actual 2002 sales greater than or equal to \$10,000 and no reported survey data from 1997 or later.

In this case $P = Q = 5$ and the calibration equations (2) become

$$\sum_{g=1}^{5} \widetilde{a}_g n_{hg} = N_h \tag{23}$$

$$\widetilde{a}_g = (p(\widehat{\beta}_g))^{-1} \tag{24}$$

where n_{hg} is the number of respondents in X-group g and Z-group h. Unless the 5×5 matrix $(n_{hg})_{g,h=1,\cdots,5}$ is singular, (23) has a unique solution $\widetilde{a}_g$, and $c_i = a_i = \widetilde{a}_g$ is the calibration weight for respondent i in X-Group g. If all the $\widetilde{a}_g$ are in the range of $p(\eta)$, a unique solution for the calibration equation (2) will be found. The fitted weights $\widetilde{a}_g$ derived from (23), a derivation that does not depend on the back-link function, are given in Table 1.

Notice that the calibration weight is less than 1 for respondents from X-group 1 in IL and for respondents from X-group 2 in LA and SD. These are the only states and X-groups where this anamoly happens, which is why we are examining them

here. If these weights were used to estimate the number of farms in one of those X-groups, the estimate would be less than the actual number of responding farms in the X-group, an absurdity. It would also imply, under our assumed quasi-randomization model, that the estimated probability of a unit in this group responding would exceed 1, also absurd.

When such an anomaly occurs, (24) has no solution using the logistic back-link function; in this case, $p(\beta_g) = (1 + exp(-\beta_g))^{-1}$ for i in X-group g. As a result, iterated use of the update equation (8) fails to converge as $\widehat{\beta}_g \to \infty$ for some g. One solution is to force the calibration weights in an offending X-group to be unity and to remove the group's observations from the sample and the calculation of the benchmark targets. When that is done, one has the calibration problem with $P > Q$.

In fact, the calibration program we have written accomplishes this automatically in a different fashion. The update equation (8) requires two matrix inversions: the first to calculate $\mathbf{W}$ and the second to invert $\widehat{\mathbf{H}}'\mathbf{W}\widehat{\mathbf{H}}$. These matrices can become numerically singular for a variety of reasons. When inverting a positive definite matrix $\mathbf{G}$, we calculate its spectral decomposition $\mathbf{G} = \sum_r \lambda_r \mathbf{e}_r \mathbf{e}_r'$, where $\lambda_1 \geq \lambda_2 \geq \cdots$ and eliminate eigenvalues/vectors for which $\lambda_r < \epsilon\lambda_1$ for some small ϵ (we use $\epsilon = .0000001$). Then we use a 'numerical generalized inverse' $\mathbf{G}^- = \sum_r \lambda_r^{-1} \mathbf{e}_r \mathbf{e}_r'$ where the sum is taken over the remaining eigenvalues/vectors. Thus for IL, LA, and SD when the X-groups are used to define model variables, we effectively end up with four model variables and five benchmark targets.

It should be noted, however, that in principle the appropriate way to approach nonresponse calibration in these states is to treat the probability of response in X-group 1 in IL (and X-group 2 in LA and SD) as if it has a known value, 1 and to treat the problem as a four model variable calibration. Otherwise, our underlying assumptions fail since otherwise neither $\mathbf{W}$ nor $\widehat{\mathbf{H}}$ is of full rank. The 'numerical generalized inverse' approach achieves the same numerical results and can be viewed as an automated method to reduce the number of model variables when that becomes necessary.

In a second set of calibration runs, we model response as a logistic function of three calibration $\mathbf{x}$-variables: an intercept, *logsales* the logarithm of the actual annual sales in 2002, truncated to the range \$1,000 to \$100,000, and $s97$ an indicator variable for whether or not the farm responded to a survey since the 1997 Census of Agriculture. For these runs, the fitted benchmark totals $\widehat{t}_{\mathbf{z}}(\widehat{\beta}))$ differ from the benchmark target totals in each state. Both sets are displayed in Table 2.

Table 2 shows how well calibration does at estimating the benchmark totals when there are less model variables than benchmark variables. This is a simple check on the appropriateness of the model. Relative to the calculated standard errors, the four-model-variable calibrations in the three problem states IL, LA, and SD appear to fit the data well. Sadly, the three-model-variable calibrations do not, especially in Illinois.

In the next section we perform a simulation study using this data. Our conclusion is that if the response model (that is, that the probability of response is the logistic back link of a linear combination of an intercept, ntercept, *logsales*, and $s97$) is correct then calibration works well. It seems that Table 2 establishes that this response model is, in fact, unlikely to be correct.

Table 2: Benchmark targets and fitted totals for three and four model variable calibrations

	N_1	N_2	N_3	N_4	N_5
CA targets	21804	14622	14309	14777	4769
CA 3 var. fit	21861.5	14578.0	14273.8	14816.0	4751.6
	(30.2)	(34.3)	(36.5)	(25.7)	(9.4)
DE targets	628	369	334	517	216
DE 3 var.fit	638.9	370.3	311.5	535.4	207.9
	(5.9)	(7.0)	(6.5)	(5.2)	(2.5)
IL targets	20220	15959	14241	24505	6513
IL 4 var. fit	20226.4	15955.6	14241.3	24505.0	6509.7
	(49.0)	(26.1)	(2.4)	(0.3)	(25.1)
IL 3 var.fit	21044.1	15597.6	13560.2	25291.4	5874.4
	(35.0)	(36.2)	(32.3)	(30.4)	(23.2)
LA targets	10390	7850	4275	2638	1040
LA 4 var. fit	10375.4	7873.3	4266.0	2639.2	1039.1
	(17.1)	(27.3)	(10.6)	(1.5)	(1.1)
LA 3 var. fit	10394.3	7880.7	4225.2	2664.8	1027.9
	(20.9)	(24.8)	(19.8)	(12.8)	(9.3)
SD targets	5847	5304	7278	11134	2416
SD 4 var. fit	5833.7	5327.0	7270.8	11135.8	2411.7
	(11.7)	(20.3)	(6.3)	(1.6)	(3.8)
SD 3 var. fit	5864.2	5322.9	7198.1	11209.4	2380.7
	(15.9)	(18.8)	(22.2)	(21.4)	(10.1)

Standard errors, calculated using (19) with $\pi_i = \pi_{ij} = 1$ and $\mathbf{y}_i = \mathbf{z_i}$, are given in parentheses.

Table 3: Estimated number of farms

	reweighting using 5 Z-groups	calibration using X-groups (4/5 variables)	calibration using intercept *logsales*, *s97*
CA	45312.5 (45.8)	46181.3 (165.6)	46178.8 (56.3)
DE	1390.9 (9.2)	1416.1 (18.7)	1400.6 (11.0)
IL	57332.4 (52.1)	60797.8 (1.5)	58925.7 (53.7)
LA	16139.6 (30.3)	16138.4 (50.8)	16425.3 (37.6)
SD	23260.7 (27.7)	23734.8 (42.1)	23821.1 (25.5)

Standard errors in parentheses.

We now use calibration to estimate the total number of active farms. In this context our y variable is a 0-1 variable for being an active farm. We compare the existing NASS approach of reweighting within Z-groups, calibration using five (for CA and DE) or four (for IL, LA, and SD) indicator variables defined by the X-groups, and calibration using the three variable model (intercept, *logsales* and *s97*). The results are given in Table 3, with the standard errors, calculated as in Section 5 for the calibration estimates and using (5) for the reweighting estimate. For the latter standard error, the standard poststratification variance formula produces the same answer. They are identical asymptotically and within roundoff error in this application.

In interpreting Table 3, it is important to realize that the standard errors are computed assuming the underlying model for that fit is correct. Thus the post-stratification standard error assumes that response probability takes on one of five possible values. The choice of one of these five possible values depends upon NASS assigned expected sales and participation in surveys since 1997.

The calibration estimate using four or five X-groups also assumes that response probability takes on one of five possible values. Here, however, actual sales is more plausibly employed in forming the groups. As a result, the estimated farm count turns out to be higher in 46 of the 47 states NASS used reweighting to adjust for census nonresponse (not shown), suggesting that the agency's nonresponse-adjustment methodology is biased downward. The high standard errors resulting in most states from employing actual sales in group formation may be the price one has to pay for a systematic reduction in this bias.

Observe that the standard error for the Illinois four-X-group fit is surprisingly small. This appears to be a ramification of assuming the probability of response in X-group 1, the only X-group containing nonfarms, is known.

The three-variable-calibration model assumes that response probability varies continuously with actual sales given survey participation. Although this seems reasonable and results in estimated standard error only slightly higher than those for poststratification, we saw in Table 2 that this model is not supported by the data. Clearly, more research on model-fitting techniques for response modelling through calibration is needed.

8 Simulations

As discussed in the previous section, an incorrect model is a possible explanation for the poor fit provided by the three variable models evidenced in Table 2. To explore this question further, we conducted simulations. Our conclusion is that, if the form of the response model is correct, then calibration performs just fine.

Our approach was as follows. For each of the five states, assuming the fitted three-variable response model is correct, response probabilities were calculated for each of the respondents to the Census. Then a synthetic state was created using only the respondents to the Census, together with their response probabilities. In particular, new target benchmarks were calculated as sizes of the Z-groups *within the synthetic state populations*, that is the respondents within the original states.

Each Monte Carlo replication consisted of creating a respondents within the synthetic states using the assigned response probabilities and fitting a calibration model to the synthetic state using these respondents. 1000 Monte Carlo simulations were done for each state. Table 4 gives for the each of the synthetic state simulations, the mean of the fitted targets and their sample standard deviation, together with square root of the mean estimated variance given by equation (19).

Examining Table 4, all mean fitted targets are within two standard errors of their benchmarks; in CA, DE, and LA they are well within one standard error. It should be noted that in CA, DE, and LA the maximum of the assigned response probabilities is 0.925, 0.859, and 0.917 respectively. By contrast, 21.5% of the synthetic IL population has an assigned probability in excess of 0.95. In SD the corresponding percentage is 23.2%. Nonlinearities in the response probability link function are most severe for probabilities close to 1 or to 0. Thus the biases in the fitted targets that possibly exist for these two populations are likely due to the large proportion of these synthetic populations with high response probabilities.

When $\mathbf{W}(\beta) = \left[\widehat{\mathbf{Var}}(\hat{\mathbf{t}}_{\mathbf{z}}(\beta))\right]^{-\mathbf{1}}$, equation (12) yields the estimate

$$\widehat{Var}(\widehat{\beta}) = \left[\widehat{\mathbf{H}}'\mathbf{W}\widehat{\mathbf{H}}\right]^{-1}. \tag{25}$$

Table 5 gives, for each synthetic state, the true β_*, the mean of the fitted $\widehat{\beta}$, their sample standard deviations, and the square root of the mean of the sample estimates calculated using (25).

Table 5 also gives the results for each state, the total number of farms $t_{\mathbf{y}}$ and its estimate $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ given by equation (15). The sample estimate of the variance of $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$ is given by equation (19).

Table 5 in general also shows good fit.

We also fit two misspecified models to the synthetic states. In the first of these, the model variables used were the five X-group variables defined in the previous section. In the second, poststratification/reweighting using the five Z-groups was used, that is the five Z-group variables were used for both the model variables and the benchmark variables.

Recall that in the synthetic states, the response probability was constructed to depend upon actual sales and participation in a previous survey. The three variable calibrations, whose results are given in Tables 4 and 5, hypothesize the correct functional form for the dependence of the response probability on these two variables

Table 4: Benchmark targets and fitted totals for three model variable calibrations, simulated populations

	N_1	N_2	N_3	N_4	N_5
CA targets	19603	13034	12502	12463	3967
CA fitted	19603.28	13033.60	12502.34	12462.83	3966.96
	28.11	32.42	34.01	23.26	8.64
	28.42	32.16	33.75	23.37	8.75
DE targets	537	311	259	435	167
DE fitted	537.02	310.99	259.04	434.96	166.99
	5.39	6.40	6.04	4.75	2.31
	5.42	6.42	5.96	4.73	2.30
IL targets	18542	14000	12048	20268	4388
IL fitted	18540.11	14001.55	12048.90	20266.67	4388.91
	31.11	31.97	27.91	25.51	20.44
	30.88	32.39	28.72	25.91	20.55
LA targets	9054	6938	3673	2182	806
LA fitted	9054.15	6938.15	3672.75	2182.09	805.87
	19.64	22.65	18.39	11.37	8.64
	19.30	22.61	18.24	11.14	8.45
SD targets	5358	4876	6499	9378	1837
SD fitted	5357.63	4876.97	6498.43	9378.30	1836.66
	14.57	17.52	20.22	19.24	9.34
	14.37	17.23	20.01	19.06	9.23

Based upon 1000 simulations. For each fitted model, the first line gives mean fitted targets, the second their empirical standard deviations, and the third the square root of the mean of the sample estimated variances calculated using (19).

Table 5: Coefficients β_* and total number of farms $t_{\mathbf{y}}$, together with estimates $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$, three model variable calibrations on simulated populations

	Coefficients			
	int	*logsales*	*s97*	Number of Farms
CA β_* and $t_{\mathbf{y}}$	3.7478	-0.2341	0.3841	39568
CA fitted $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$	3.7433	-0.2335	0.3835	39565.53
	0.1051	0.0132	0.0510	52.81
	0.1062	0.0132	0.0507	53.57
DE β_* and $t_{\mathbf{y}}$	2.3161	-0.0964	0.1543	1147
DE fitted $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$	2.3292	-0.0957	0.1379	1146.61
	0.4934	0.0638	0.2227	10.63
	0.4908	0.0628	0.2230	10.16
IL β_* and $t_{\mathbf{y}}$	5.2025	-0.4548	1.1894	48608
IL fitted $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$	5.2009	-0.4545	1.1886	48605.32
	0.1193	0.0152	0.0525	48.59
	0.1207	0.0154	0.0540	48.30
LA β_* and $t_{\mathbf{y}}$	3.5327	-0.2617	0.6822	13944
LA fitted $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$	3.5277	-0.2609	0.6813	13944.63
	0.1899	0.0269	0.0958	35.56
	0.1898	0.0269	0.0952	35.21
SD β_* and $t_{\mathbf{y}}$	6.0846	-0.5188	1.2285	20231
SD fitted $\widehat{\beta}$ and $\widehat{t}_{\mathbf{y}}(\widehat{\beta})$	6.0915	-0.5193	1.2281	20230.13
	0.2287	0.0272	0.0846	24.20
	0.2316	0.0276	0.0856	23.95

Based upon 1000 simulations. For each fitted model, the first line gives mean fitted values, the second their empirical standard deviations, and the third the square root of the mean of the sample estimated variances calculated using (25) or (19).

Table 6: Estimated number of farms, simulated populations

	Reweighting: 5 Z-groups			Calibration: 4/5 X-variables		
		standard error			standard error	
No. farms	mean	empirical	eqn (19)	mean	empirical	eqn (19)
CA 39568	38815.03	39.62	42.86	39365.00	154.60	156.68
DE 1147	1138.96	8.01	8.41	1145.66	15.97	16.91
IL 48608	47369.47	37.58	45.38	48133.22	73.04	79.66
LA 13944	13708.36	26.03	27.92	13713.93	81.04	91.46
SD 20231	19766.06	19.51	25.17	20081.41	43.86	51.50

Based upon 1000 simulations.

(but fit the coefficients). The five X-group calibrations use the correct variables to model the response probability but implicitly hypothesize an incorrect functional form for its dependence upon these variables. By contrast, poststratification models response probability using the wrong variables (NASS expected sales instead of actual sales).

Another difference between the five X-group variable calibration fit and the poststratification fit is that in the former case, the unique solution to equation (23) might not satisfy equation (24). That is, the unique solution to (23) corresponds to response probabilities outside the range $(0, 1)$. This leads to the fitted targets not matching the preset targets. As discussed in the previous section, this anomaly can be handled by setting the response probabilities in the offending X-group to one, or equivalently, by the variable-dropping technique described there. This anomaly occurred for 0%, 26.3%, 89.3%, 24.5%, and 48.8% of the runs in CA, DE, IL, LA, and SD respectively. This anomaly is mathematically impossible with poststratification.

For these two response models, Table 6 gives the total number of farms, and the mean of the sample estimates, their empirical standard deviation, and the square root of the mean of the sample estimates of their variance as calculated using equation (19). Examining these two misspecified models, it is clear that although both of them produced downward biased estimates, the bias is in most cases substantially reduced when the appropriate variables are used for modelling the response mechanism. This provides a strong argument for separating the calibration variables from the response model variables.

9 Concluding Remarks

As noted in the introduction, it is a simple matter to adapt most of the results in this paper to the situation where calibration is used to adjust for coverage errors, whether from frame undercoverage or unit duplication. In this context, p_i is the expected number of times population unit i is in the frame. Under the assumed quasi-randomization model this probability is independent of the actual sample drawn and of how many times other elements of the population are in the frame.

The possibility a unit appearing more than once on a frame does cause some small differences in variance estimation. When there is no possibility of unit duplication, the variance of the function indicating whether population unit i is in the frame is $p(\mathbf{x}_i^{'}\beta_*)(1-p(\mathbf{x}_i^{'}\beta_*))$, paralleling the situation with survey response. When the frame contains potential duplication, however, the variance of the expected number of times a population unit is no longer $p(\mathbf{x}_i^{'}\beta_*)(1-p(\mathbf{x}_i^{'}\beta_*))$ and its value will need to be assumed for the analogue of (18), a measure of the added variance due to frame errors, to be derived.

In the section 7, we described how NASS used poststratification, a special case of calibration, to adjust for nonresponse in the 2002 Census of Agriculture. NASS also used calibration to adjust for undercoverage. It did this by treating the respondent sample as the original survey sample with "sample weights" equal to the nonresponse-adjustment weights. A single set of variables served as both the model and benchmark variables. The population totals for some benchmarks were wholly derived from independent NASS surveys. For other benchmark variables, nonresponse-adjusted Census aggregates were added to estimates from an independent area sample of farms not on the Census frame.

Obviously, there is a lot of improvement possible in this methodology as there is in the nonresponse adjustment. The main point we want to make here, however, is that although the Census of Agriculture used calibration techniques to adjust for both nonresponse and undercoverage, these adjustments were done separately. In fact, some of the coverage benchmarks could only be determined after the nonresponse adjustment.

Suppose one wanted to calibrate simultaneously for nonresponse and coverage errors. Let $\mathbf{x}_1$ be the vector of variables to be used to model coverage and $\mathbf{x}_2$ the vector to model response. Let β_1 and β_2 be the corresponding coefficients. If we let $\mathbf{x}=[\mathbf{x}_1^{'} \quad \mathbf{x}_2^{'}]^{'}$, $\beta=[\beta_1^{'} \quad \beta_2^{'}]^{'}$, and $p(\mathbf{x},\beta)=p_1(\mathbf{x}_1^{'}\beta_1)p_2(\mathbf{x}_2^{'}\beta_2)$, then most of our results can be extended with little effort.

It is important in this setup that the variables used to model response be different from those used to model coverage. Indeed the pair $(\mathbf{x}_1,\mathbf{x}_2)$ should not be close to collinear. Otherwise

$$\begin{aligned} p_1(\mathbf{x}_1^{'}\beta_1)p_2(\mathbf{x}_2^{'}\beta_2) &\approx p_1(\mathbf{x}_1^{'}\beta_{1*})p_2(\mathbf{x}_2^{'}\beta_{2*}) \\ &+ p_1(\mathbf{x}_1^{'}\beta_{1*})p_2^{'}(\mathbf{x}_2^{'}\beta_{2*})\mathbf{x}_2^{'}(\beta_2-\beta_{2*}) \\ &+ p_1^{'}(\mathbf{x}_1^{'}\beta_{1*})p_2(\mathbf{x}_2^{'}\beta_{2*})\mathbf{x}_1^{'}(\beta_1-\beta_{1*}) \end{aligned}$$

As a result, (β_1,β_2) cannot be nearly estimated with a first-order approximation. In practice what would happen will be a breakdown of the asymptotic approximations used here as well as slow numerical convergence in the calculation of $(\widehat{\beta}_1,\widehat{\beta}_2)$.

An example where response and coverage calibration has been done simultaneously can be found in Crouse and Kott (2004). Rather than assuming the components of the response and coverage models are distinct, however, that paper supposes that the components are, or could be, exactly the same. This is accomplished by assuming both models have the form $p(.)=exp(.)$ so that $p_i=p_1(\mathbf{x}_{i1}^{'}\beta_1)p_2(\mathbf{x}_{i2}^{'}\beta_2)$ is also of that form. The adjustments for nonresponse and coverage errors can not separated, but as the real goal of calibration is the estimation of totals (or functions of totals), that need not be considered a problem.

Much work is needed in determining how to select model and benchmark variables in practice, especially when the targets of the benchmark variables are themselves potentially subject to sampling and measurement errors. Nevertheless, by allowing a separation between the model and benchmark variables, the approach to calibration developed here may open the door to more plausible modelling of the response and coverage mechanisms.

Avoiding response and coverage modelling as much as possible remains a prudent policy. Even the cleverest model assumptions are difficult to test and prone to failure. Unfortunately, eschewing models will not be viable option as surveys based on incomplete frames or suffering from small response rates become increasingly common.

References

Crouse, C. and Kott, P.S. (2004), "Evaluating Alternative Calibration Schemes for an Economic Survey with Large Nonresponse," *ASA Proceedings of the Survey Reseach Methods Section*.

Folsom., R.E. (1991). "Exponential and Logistic Weight Adjustment for Sampling and Nonresponse Error Reduction," *ASA Proceedings of the Social Statistics Section*, 197-202.

Folsom., R.E. and Singh, A.C. (2000), "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Post-stratification," *ASA Proceedings of the Section on Survey Research Methods*, 598-603.

Fuller, W.A. and Isaki, C.T. (1981), "Survey Design Under Superpopulation Models," in *Current Topics in Survey Sampling*, eds. D. Krewshi, J.N.K. Rao, and R. Platek, New York: Academic Press.

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.

Kim, J.K., Navarro, A., and Fuller, W.A. (2006), "Replicate Variance Estimation for Two-Phase Stratified Sampling," *Journal of the American Statistical Association*, forthcoming.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Kott, P.S. (2004a), "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors," Paper presented at Mini Meeting on Current Trends in Survey Sampling and Official Statistics, Calcutta, India, http://www.nass.usda.gov/research /reports/ calcutta-for-web.pdf.

Kott, P.S. (2004b), "Collected Technical Notes on Weighting and its Impact on the 2002 Census of Agriculture," http://www.nass.usda.gov/research/reports/ COMBINED.pdf.

Lohr, S. L. (1999), *Sampling Design and Analysis*, Pacific Grove, CA: Duxbury Press.

Lundström, S. and Särndal, C.-E. (1999), "Calibration as a Standard Method for the Treatment of Nonresponse," *Journal of Official Statistics*, 15, 305-327.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second edition, London: Chapman & Hall.

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.

Thompson, M. E. (1997), *Theory of Sample Surveys*, London: Chapman & Hall.