Can Calibration Be Used to Adjust for "Nonignorable" Nonresponse?

Phillip S. Kott and Ted Chang

Calibration can be used to adjust for unit nonresponse when the model variables on which the response/nonresponse mechanism depends do not coincide with the benchmark variables in the calibration equation. As a result, model-variable values need only known for the respondents. This allows the treatment of what is usually considered nonignorable nonresponse. Although one can invoke either quasi-randomization or prediction-model-based theory to justify the calibration, both frameworks rely on unverifiable model assumptions, and both require large samples to produce nearly unbiased estimators even when those assumptions hold. We will explore these issues theoretically and with an empirical study.

KEY WORDS: Prediction model; Quasi-randomization; Benchmark variable; Model variable; Bias; Response-drive response group.

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030, pkott@nass.usda.gov Ted Chang is Professor, Department of Statistics, University of Virginia, Charlottesville VA 22904, tcc8v@virginia.edu.

1. Introduction

Although not orginally designed for that purpose, calibration can be used to adjust for unit nonresponse. See, for example, Kott (2006). It is less well known that calibration can be employed when the (explanatory) model variables on which the response/nonresponse mechanism depends do not coincide with the benchmark variables in the calibration equation. As a result, model-variable values need only be known for the respondents. This allows the treatment of what is usually considered nonignorable nonresponse.

Section 2 lays out the two theories behind using calibration as a method for nonresponse adjustment: quasi-random response modeling and prediction modeling. Section 3 extends the prediction-modeling approach to cover nonignorable nonresponse. The response mechanism is said to be "nonignorable" when the expected value of the survey variable under the prediction model differs between respondents and nonrespondents even when conditioned on benchmark variables.

Only the prediction-modeling approach needs to be extended to cover nonignorable nonresponse. This is because the ignorability of the response mechanism is an irrelevant concept under quasi-random response modeling since the unit propensities of response are modeled in that approach, not the survey variable.

A version of the prediction model in the approach bearing its name relates the survey variable to the model variables. In the extension, a second model equation, called the "measurement-error model," connects the model variables to the benchmark variables.

The respective theories behind the quasi-random-response and prediction-modeling approaches rely on samples being large and on model assumptions that can fail in practice. We explore this empirically in Sections 4 and 5 for a census, thereby avoiding the added complication of a random-sampling component in the estimates.

Mutually exclusive group-indicator variables known for all units in the population serve as the benchmark variables in our empirical evaluations. The "benchmark groups" themselves are based on previously-collected frame information. Following Kott (2005), model variables are created by constructing analogous "model groups" using survey information known only for the respondents.

Neither the prediction/measurement-error model nor the response model justifying calibration using these model groups is correct. Both, however, are closer to the truth than the models justifying calibration treating the benchmark groups as the model groups. As a consequence, using the response-generated model groups leads to much lower empirical biases and smaller mean squared errors if slightly larger empirical standard deviations.

We had hoped that a prediction-model-based correction to the quasi-randomization mean-squared-error estimate would prove to be effective even when the models supporting the prediction-model approach were not strictly true. Unfortunately, this turns out not to be the case even when the model variables in the calibration are based on the true quasi-random response model.

Section 6 provides a discussion of these and some additional empirical results. In particular, we show there can be efficiency gains from a two-step calibration under a simple random sample with unit nonresponse. This approach first poststatifies the sample using the benchmark groups and then creates and employs analogous model groups to adjust for the nonresponse.

2. Some theory

2.1. Calibration

Linear calibration weights can be put in the form:

$$w_k = d_k(1 + \mathbf{z}_k^T \mathbf{g}), \tag{1}$$

where k denotes an element in population U,
$\{d_k\}$ is the set of original sample weights – the inverses of the element selection probabilities (for a census, all $d_k = 1$) – for the elements in sample S,
$\mathbf{z}_k = (z_{k1}, ..., z_{kP})^T$ is a P-vector with $z_{k1} = 1$ (or the equivalent: $\mathbf{z}_k^T \lambda = 1$ for some λ),
$\mathbf{g} = (\sum_{j\in S} d_j \mathbf{x}_j \mathbf{z}_j^T)^{-1}(\sum_{j\in U} \mathbf{x}_j - \sum_{j\in S} d_j \mathbf{x}_j)$, and
$\mathbf{x}_k$ is a P-vector of benchmark (or calibration) variables for which $\sum_{j\in U} \mathbf{x}_j$ is known.
For convenience, we will assume that matrices such as $N^{-1}\sum_{j\in S} d_j \mathbf{x}_j \mathbf{z}_j^T$ when encountered in the theoretical sections of this paper are of full rank.

The weights in equation (1) are constructed so that the calibration equation,

$$N^{-1}\sum_{k\in S} w_k\mathbf{x}_k = N^{-1}\sum_{k\in U} \mathbf{x}_k, \tag{2}$$

holds. In most multivariate applications, the vector $\mathbf{z}_k$ coincides with $\mathbf{x}_k$, but that will not generally be the case here. In linear regression, the components of $\mathbf{z}_k$ when not equal to corresponding components of $\mathbf{x}_k$ are called "instrumental variables."

Under mild conditions which we assume to hold, $t = \sum_{k\in S} w_k y_k$ is a randomization-consistent estimator for $T = \sum_{k\in U} y_k$. In addition, t is an unbiased predictor for T under the linear prediction model:

$$y_k = \mathbf{x}_k^T\beta + \varepsilon_k, \tag{3}$$

for each $k \in U$, where

$$E\{\varepsilon_k |(\mathbf{x}_j, \mathbf{z}_j, I_j ; j\in U)\} = 0, \tag{4}$$

and $I_j = 1$ when $j \in S$, 0 otherwise.

2.2 Nonresponse

Linear calibration can also be used to adjust for nonresponse. Redefining S as the subset of sample respondents, t remains a prediction-model unbiased predictor for T. Equation (4) requires that the prediction model in equation (3) hold whether or not k is in S. This now means the response mechanism, like the sampling mechanism, is ignorable.

Strictly speaking, the sampling and response mechanism is ignorable if the *distribution* of $\varepsilon_k |(\mathbf{x}_j, \mathbf{z}_j, j\in U)$ is the same regardless of the I_j. For most practical purposes, however, it suffices to focus on the conditional expectation of the ε_k.

An alternative justification for linear calibration as a method of nonresponse adjustment treats sample response as an additional phase of probability sampling. Each element in the population is assumed to have a Poisson probability of response equal to

$$p_k = \rho(\mathbf{z}_k^T\gamma) = 1/(1 + \mathbf{z}_k^T\gamma). \tag{5}$$

If $\mathbf{g} \approx \gamma$, then $t = \sum_{k \in S} w_k y_k \approx \sum_{k \in S} d_k(1 + \mathbf{z}_k^T \gamma) y_k$ would be randomization consistent under mild conditions. This was first noted by Fuller *et al*. (1994) for the $\mathbf{z}_k = \mathbf{x}_k$ case.

The form of the response model in equation (5) is unlikely, but it may be reasonable to assume a response (propensity) model of the form

$$p_k = \rho(\mathbf{z}_k^T \gamma) = 1/f(\mathbf{z}_k^T \gamma) \tag{6}$$

where $f(\delta)$ is an appropriately-chosen monotonic and twice differentiable function. Subject to a set of mild conditions, if a vector $\mathbf{g} \approx \gamma$ can be found that satistifes

$$\sum_{k \in S} d_k f(\mathbf{z}_k^T \mathbf{g}) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k, \tag{7}$$

then using the calibration weights, $w_k = d_k f(\mathbf{z}_k^T \mathbf{g})$, results in a randomization consistent estimator under the response model.

Folsom and Singh (2000) describe an iterative method for finding such a $\mathbf{g}$ when $\mathbf{z}_k = \mathbf{x}_k$. Kott (2006) provides the obvious extension to a more general $\mathbf{z}_k$. This extension allows the explantory variables in the response model to differ from the benchmark variables in the calibration equation. Earlier work along this line in France is described by Sautory (2003).

Chang and Kott (2007) call $f(\delta)$ the "back-link function" because it is the inverse transformation of the link function in the generalized-linear-model literature. See, for example, see McCullagh and Nelder (1983). They also call the components of $\mathbf{z}_k$ "model variables."

When a $\mathbf{g}$ satisfying equation (7) can be found, $t = \sum_{k \in S} w_k y_k = \sum_{k \in S} d_k f(\mathbf{z}_k^T \mathbf{g}) y_k$ is both quasi-randomization consistent (randomization consistent under the response model) and prediction-model unbiased for T under mild conditions. The former property is defined with respect to the model in equation (6) and the latter to the model in equation (3). Observe that *only one of the two models need hold for t to be a nearly unbiased estimator for T in some sense.* This property has been called "double protection" or "double robustness" in the biometrics literature (see, for example, Bang and Robbins, 2005).

Note that if the back-link function in the response model is incorrect, then **g** defined implicitly by the equation (7) bares no relationship to γ as defined in equation (6). In fact, from a response-model-free point of view, γ is simply the limit of **g** as the sample size grows arbitrarily large. Some may argue that taking such a limit assumes a quasi-randomization framework. Perhaps, but the actual Poisson response probabilities need not conform to a known back-link function $f(\delta)$.

2.3. An example

The ratio estimator provides an example of a estimator that is (nearly) unbiased when either the response or prediction model holds even though the other may fail. It is

$$t = \sum_{k\in U} x_k \frac{\sum_{k\in S} d_k y_k}{\sum_{k\in S} d_k y_k}$$

$$= \sum_{k\in U} x_k \frac{\sum_{k\in S} d_k z_k y_k}{\sum_{k\in S} d_k z_k y_k}, \text{ where } z_k = 1.$$

This estimator is nearly unbiased under the response model where every element in the original sample is equally like to respond (since all $\mathbf{z}_k = z_k = 1$) *regardless of how y_k and x_k are related.* Alternatively, it is unbiased under the prediction model where $E(y_k|x_k) = \beta x_k$ (since $\mathbf{x}_k = x_k$) for both respondents and nonrespondents *even when response probabilities vary with x_k*.

2.4. Variance/Mean-squared-error Estimation

If the ε_k in the prediction model of equation (3) are uncorrelated, and $E(\varepsilon_k^2) = \sigma_k^2$, then the model variance of *t* (as a predictor for *T*) is approximately $\sum_{k\in S}(w_k^2 - w_k)\sigma_k^2$. The identity is exact when $\sigma_k^2 = \mathbf{x}_k{}^T\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda}$.

For simplicity, we will assume here that Poisson element sampling was used to draw the orginal sample or that the original sample was a census as will be the case in the empirical example in Section 4. A nearly unbiased estimator for both the prediction-model variance and quasi-randomization mean squared error of *t* (under the respective models) would be

$$v = \sum_{k \in S} (w_k^2 - w_k) e_k^2, \tag{8}$$

where

$$e_k = y_k - \mathbf{x}_k^T \{ \sum_{j \in S} d_j f_1(\mathbf{z}_j^T \mathbf{g}) \mathbf{z}_j \mathbf{x}_j^T \}^{-1} \sum_{j \in S} d_j f_1(\mathbf{z}_j^T \mathbf{g}) \mathbf{z}_j y_j, \tag{9}$$

and $f_1(\delta)$ is the first derivative of $f(\delta)$. The $f_1(\mathbf{z}_j^T \mathbf{g})$ terms in the definition of the e_k assure the nearly unbiasedness of v as an estimator for the quasi-randomization mean squared error of t. They are no more than arbitrary constants from the prediction-model point of view.

The relative (prediction-model) bias of v as an estimator of the prediction-model variance of t is $O(1/n)$ under mild conditions, while the relative bias of v as an estimator of quasi-randomization mean squared error of t is $O_P(1/n^{1/2})$. Nevertheless, it is troubling that this prediction-model bias is positive when each $E(e_k^2) < \sigma_k^2$, as will almost always be the case.

Kott and Brewer (2001) describe a number of ways to sharpen the estimation of or for prediction-model variance. One method replaces e_k^2 in equation (8) with

$$r_k^2 = \frac{e_k^2}{1 - \mathbf{x}_k^T \left\{ \sum_{j \in S} d_j f_1(\mathbf{z}_j^T \mathbf{g}) \mathbf{z}_j \mathbf{x}_j^T \right\}^{-1} d_k f_1(\mathbf{z}_k^T \mathbf{g}) \mathbf{z}_k}. \tag{10}$$

Only on rare ocassion will this procedure remove the entire prediction-model bias of v, but it will usually remove most of it, and the bias of the resulting prediction-model variance estimator will have an ambiguous sign.

3. Nonignorable nonresponse

From a prediction-model point of view, when the variables in $\mathbf{z}_k$ are *not* all deterministic functions of the $\mathbf{x}_k$, it may be desirable to replace the requirement in equation (4) with a weaker variant:

$$E\{\varepsilon_k | (\mathbf{z}_j, I_j ; j \in U)\} = 0. \tag{4'}$$

Removing the conditioning on the $\mathbf{x}_j$ in the above equation allows the possibility that the response mechanism is nonignorable under the linear prediction model, $y_k = \mathbf{x}_k^T\boldsymbol{\beta} + \varepsilon_k$, since the expectation of ε_k need not be zero when conditioned on $\mathbf{x}_j$, I_j ; $j \in U$. In particular, $E\{\varepsilon_k | (\mathbf{x}_j ; j \in U)\}$, although zero on average when the model holds for the population, may differ between respondents and nonrespondents.

Some care is necessary when equation (4') replaces (4), because $\mathbf{g}$ in $w_k = d_k f(\mathbf{z}_k^T\mathbf{g})$ is a function of the sampled $\mathbf{x}_j$. As a result, the strict prediction-model unbiasedness of t becomes near unbiasedness under mild assumptions. One such assumption is that the limit of $\mathbf{g}$ exists as the sample size grows arbitrarily large. Moreover, this limit, again called "$\boldsymbol{\gamma}$," is assumed not to be a function of the sampled $\mathbf{x}_j$.

As a result, even when the second equation in (6) does not hold, so long as $\mathbf{g} = \boldsymbol{\gamma} + \mathbf{O}_P(1/n^{1/2})$, we have

$$
\begin{aligned}
E[N^{-1}(t - T) \,|\{ \mathbf{z}_j, I_j ; j \in U\}] &= E\{N^{-1}\sum_{k \in S} w_k \varepsilon_k | (\mathbf{z}_j, I_j ; j \in U)\} \\
&= E\{N^{-1}\sum_{k \in S} d_k f(\mathbf{z}_k^T\mathbf{g})\varepsilon_k | (\mathbf{z}_j, I_j ; j \in U)\} \\
&= E\{N^{-1}\sum_{k \in S} d_k f(\mathbf{z}_k^T\boldsymbol{\gamma})\varepsilon_k | (\mathbf{z}_j, I_j ; j \in U)\} + O_P(1/n^{1/2}) = O_P(1/n^{1/2}).
\end{aligned}
$$

An example of a framework in which equation (4') is sensible follows. Suppose the y_k can be fit by this prediction model:

$$y_k = \mathbf{z}_k^T\boldsymbol{\theta} + \tau_k, \tag{11}$$

where $E\{\tau_k | (\mathbf{z}_j, I_j ; j \in U)\} = 0$. In addition, suppose the benchmark variables can be fit by a "measurement-error" model:

$$\mathbf{x}_k^T = \mathbf{z}_k^T\boldsymbol{\Gamma} + \boldsymbol{\xi}_k^T, \tag{12}$$

where $E\{\boldsymbol{\xi}_k | (\mathbf{z}_j, I_j ; j \in U)\} = \mathbf{0}$ ("measurement-error" is in quotes because this use of measurement-error modeling is idiosyncratic). It is not hard to see from equations (11) and (12) that $\boldsymbol{\beta}$ in equation (3) is $\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}$, while ε_k is $\tau_k - \boldsymbol{\xi}_k^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}$. If the $(\tau_k, \boldsymbol{\xi}_k^T)^T$ are

uncorrelated across the k, then so are the ε_k. This is handy for prediction-model variance estimation as described in Section 2 and modified to allow the assumption in equation (4') to replace that in (4).

Both the prediction model in equation (11) and measurement-error model in (12) assume the response and sampling mechanisms are ignorable *conditioned on* $\mathbf{z}_k$ rather than $\mathbf{x}_k$. Indeed, the components of $\mathbf{z}_k$ have become model (explanatory) variables just as in the quasi-randomization framework.

Equation (12) need *not* be a causal model. Indeed, the components of $\mathbf{x}_k$ are often frame variables determined before the sample is enumerated, while the components of $\mathbf{z}_k$ can include survey values, perhaps even y_k itself. It is important to remember that our goal is to estimate T in a nearly unbiased fashion. It is not to estimate θ, $\mathbf{\Gamma}$, or $\beta = \mathbf{\Gamma}^{-1}\theta$. The estimation of model parameters is, at most, a means to an end.

Chang and Kott (2007) extend the quasi-randomization approch to calibration for nonresponse to situations where there are $Q < P$ components of $\mathbf{z}_k$. They show that under mild conditions finding a $\mathbf{g}$ that minimizes the objective function:

$$\mathbf{S} = N^{-2}\left\{\sum_{k\in U}\mathbf{x}_k^T - \sum_{k\in S} d_k f(\mathbf{z}_k^T\mathbf{g})\mathbf{x}_k^T\right\}\mathbf{\Lambda}^{-1}\left\{\sum_{k\in U}\mathbf{x}_k - \sum_{k\in S} d_k f(\mathbf{z}_k^T\mathbf{g})\mathbf{x}_k\right\},$$

where $\mathbf{\Lambda}$ is a positive-definite P x P matrix, will produce a randomization consistent $t = \sum_{k\in S} w_k y_k = \sum_{k\in S} d_k f(\mathbf{z}_k^T\mathbf{g})y_k$ under the response model in equation (6).

For a given $\mathbf{\Lambda}$, this means finding a $\mathbf{g}$ such that

$$N^{-2}\left\{\sum_{k\in U}\mathbf{x}_k^T - \sum_{k\in S} d_k f(\mathbf{z}_k^T\mathbf{g})\mathbf{x}_k^T\right\}\mathbf{\Lambda}^{-1}\sum_{j\in S} d_j f_1(\mathbf{z}_j^T\mathbf{g})\mathbf{x}_j\mathbf{z}_j^T = \mathbf{0}, \qquad (13)$$

In some sense, the optimal choice for $\mathbf{\Lambda}$ is the quasi-randomization variance of $\mathbf{Q} = (n^{1/2}/N)\sum_{j\in S} d_j f(\mathbf{z}_j^T\gamma)\mathbf{x}_j$, but that variance cannot be estimated directly because γ is unknown. As a result, Chang and Kott suggest replacing $\mathbf{\Lambda}$ in equation (13) with $\hat{\mathbf{\Lambda}}(\mathbf{g})$, an estimator for the randomization variance of $\mathbf{Q}$ under the assumption that

$\gamma = \mathbf{g}$. This value, which is $\hat{\mathbf{\Lambda}}(\mathbf{g}) = (n/N^2)\sum_{j\in S}[\{d_j f(\mathbf{z}_j{}^T\mathbf{g})\}^2 - d_j f(\mathbf{z}_j{}^T\mathbf{g})]\mathbf{x}_j\mathbf{x}_j{}^T$ in our special case, gets revised in each iteration of the process used to find $\mathbf{g}$.

From a prediction-model viewpoint, given a P x Q matrix $\boldsymbol{A}$ of full rank, we can transform the calibration equation in equation (2) into $N^{-1}\sum_{k\in S} w_k\mathbf{x}_k{}^T\boldsymbol{A} = N^{-1}\sum_{k\in U}\mathbf{x}_k{}^T\boldsymbol{A}$. In addition, as long as the matrix $\boldsymbol{A}$ is not a function of the sampled $\mathbf{x}_j$, we can replace $\mathbf{x}_k{}^T$ in equation (13) with

$$\mathbf{x}_k{}^T\boldsymbol{A} = \mathbf{z}_k{}^T\mathbf{\Gamma}\boldsymbol{A} + \boldsymbol{\xi}_k{}^T\boldsymbol{A}$$

or

$$\tilde{\mathbf{x}}_k{}^T = \mathbf{z}_k{}^T\tilde{\mathbf{\Gamma}} + \tilde{\boldsymbol{\xi}}_k{}^T. \tag{14}$$

The results of the prediction-model analysis then follow with $\tilde{\mathbf{x}}_k$ replacing $\mathbf{x}_k$, and $N^{-1}\sum_{k\in S} w_k\tilde{\mathbf{x}}_k = N^{-1}\sum_{k\in U}\tilde{\mathbf{x}}_k$ replacing the calibration equation.

Effectively, Chang and Kott set

$$\hat{\boldsymbol{A}} = N^{-1}\mathbf{\Lambda}^{-1}\sum_{j\in S} d_j f_1(\mathbf{z}_j{}^T\mathbf{g})\mathbf{x}_j\mathbf{z}_j{}^T, \tag{15a}$$

or

$$\begin{aligned}\hat{\boldsymbol{A}} &= N^{-1}\left\{\hat{\mathbf{\Lambda}}(\mathbf{g})\right\}^{-1}\sum_{j\in S} d_j f_1(\mathbf{z}_j{}^T\mathbf{g})\mathbf{x}_j\mathbf{z}_j{}^T \\ &= \tfrac{N}{n}\left(\sum_{j\in S}[\{d_j f(\mathbf{z}_j{}^T\mathbf{g})\}^2 - d_j f(\mathbf{z}_j{}^T\mathbf{g})]\mathbf{x}_j\mathbf{x}_j{}^T\right)^{-1}\sum_{j\in S} d_j f_1(\mathbf{z}_j{}^T\mathbf{g})\mathbf{x}_j\mathbf{z}_j{}^T\end{aligned} \tag{15b}$$

and $\hat{\mathbf{x}}_k{}^T = \mathbf{x}_k{}^T\hat{\boldsymbol{A}}$. From this perspective, equation (15) simply restates the calibration equation as $N^{-1}\sum_{k\in S} w_k\mathbf{x}_k{}^T\hat{\boldsymbol{A}} = N^{-1}\sum_{k\in U}\mathbf{x}_k{}^T\hat{\boldsymbol{A}}$.

Either version of $\hat{\boldsymbol{A}}$ in equation (15) is a function of the sampled $\mathbf{x}_j$ violating an assumption needed from a prediction/measurement-error viewpoint to transform $E\{\boldsymbol{\xi}_k\,|(\mathbf{z}_j, I_j\,; j\in U)\} = \mathbf{0}$ into the analogous $E\{\tilde{\boldsymbol{\xi}}_k\,|\{\mathbf{z}_j, I_j\,; j\in U)\} = \mathbf{0}$. We can get around this problem by letting $\boldsymbol{A}$ be the asymptotic limit of $\hat{\boldsymbol{A}}$, which we assume to exist whether or not the back-link function in equation (6) is specified correctly. Moreover, this limit is assumed *not* to be a function of the sampled $\mathbf{x}_j$. Consequently, with some

work we can establish that t is nearly prediction/measurement-error model unbiased for T under mild conditions. In addition, after $\mathbf{x}_k^T$ is replaced by $\hat{\mathbf{x}}_k^T = \mathbf{x}_k^T \hat{\boldsymbol{A}}$, when defining the e_k in equation (9), the prediction model variance of t can be estimated by v in equation (8). Finally, although analogues to the sharpened version of prediction/measurement-error model-variance (or mean-squared-error) estimation described in and around equations (10) does not provide exact unbiasedness, it should lead to improvement over v.

4. Setting up an empirical exploration

In this section, we create data sets from respondent data for the 2002 Census of Agriculture in South Dakota. We will be interested in estimating total sales from these data sets. When the data come from a census, the original sampling weight, d_k, is 1 for all k.

We will treat the entire South Dakota respondent data set with some additional created variables (and one extreme outlier removed) as a population of interest. To explore sample-size issues, we will also treat a 20% random subsample of this respondent set as a population of interest. We will refer to the two as the *100% population* and the *20% population*.

In Chang and Kott (2008), we estimated the following element response function for South Dakota:

$$p_k = \{1 + \exp(-6.085 + 0.5188\, logtvp2_k - 1.2285\, s97_k)\}^{-1}, \qquad (16)$$

where $logtvp2_k$ is the (natural) logarithm of element sales (in dollars) truncated to the range [1000, 100000], and $s97_k$ is an indicator of whether or not the element responded to any surveys within the last five years.

Equation (16) is a classic example of a nonignorable response mechanism in that response is a function of what we want to measure: sales. We will use a truncated variant of this equation to generate independent subsets of respondents:

$$p_k = 0.001 + 0.998\{1 + \exp(-6.085 + 0.5188\, logtvp2_k - 1.2285\, s97_k)\}^{-1}. \qquad (17)$$

This restricts the range of the response probabilities between 0.001 and 0.999.

Sales are not known before enumeration. That is why NASS used poststratification to adjust for nonresponse in the 2002 Census of Agriculture based on groups formed using a before-the-fact projection of sales, called "frame sales" (because it is known for all elements on the frame) and the *s97* indicator.

Unfortunately, the element frame sales have not been retained on Census data sets. Consequently, we use the actual sales reported from the 2002 Census as the *frame sales* for our two *populations* (*100%* and *20%*) and generate the *true sales* for each element using the formula:

$$\textit{true sales}_k = [.5 + \{1 + \exp(\sigma Z_k)\}^{-1}]\ \textit{frame sales}_k, \qquad (18)$$

where Z_k is a random draw from a N(0, 1) distributions, and $\sigma = 1$. Actual sales were often not recorded when less than 1000 because the element was deemed not to be a farm and therefore out of scope. When that happened, we replace the missing value with a draw from the uniform distribution on [0, 1000).

We generate five sets of *true sales* for each element in a *population*. This gives us five *100% population sets* and five *20% population sets*. Using equation (17), we generate 1000 *respondent subsets* for each of the 10 *population sets*.

We use *frame sales* to create *true sales* because the former, which are in practice often based on previously reported sales data, exists before the latter. Thus, this is the more reasonable direction for the causality despite the measurement-error model assumed in equation (12). Models in survey sampling are often little more than useful fictions. One should always keep that in mind.

We estimate total sales from each *respondent subsample* in basically two different ways. One way employs simple poststratification, the most commonly used method in survey practice. The population is divided into 10 mutually exclusive *response groups* by cross-classifying five size classes based on *frame sales* (having cut points at 1000, 10000, 50000, and 250000) with the two realizations of *s97*. In our notation, the vector of benchmark variable, $\mathbf{x}_k$ has ten components, each being a 0/1 indicator of membership in one of the groups. Poststratification assumes the model variables, the components of $\mathbf{z}_k$, are the same as the components of $\mathbf{x}_k$. The choice of the back-link function, $f(\delta)$ has no effect as long as $1/f(\delta)$, is free to attain

the realized values of group response rates; see equation (6). For simplicity, we can (usually) set $f(\delta) = \delta$. (We will explain the parenthetical limitation later in the section.)

The quasi-random response model supporting this methodolgy is that every element in a particular "benchmark group" has an equal probability of response. That is not true with our data, since response is generated using equation (17). The prediction model is that the *true sales* for every element in a particular benchmark group has the same expected value whether or not the element responds. That is also not true, since we know from equation (17) that true sales and response are strongly related, and two elements in the same benchmark group can have farly divergent *true-sales* values (see equation (18)).

In these evaluations, we know the true response model in equation (17), both its functional form and its arguments. In real life, neither is likely to be the case. For that reason, we now consider a vector of model variables, $\mathbf{z}_k$, the components of which are defined analogously to the components of $\mathbf{x}_k$, but with *true sales* replacing *frame sales*. This can be viewed as a useful locally-linear approximation for the true response model in which the probability of response is the same for any two elements in the same *model group*, that is to say, two elements having the same value for each component of $\mathbf{z}_k$.

The prediction model supporting this calibration stipulates is that each element in the same model group has the same expected value of true sales regardless of whether or not it responds. Not quite true – since response is correlated with true sales, but close – since true sales do not vary by much within a model group. The measurment error model is that the components of $\mathbf{x}_k$ of two elements in the same model group have the same expected value whether or not they respond. It is likewise close to being true.

To perform the calibration, we would like $f(\delta)$ again to be δ. If that were so, we would have $w_k = d_k(1 + \mathbf{z}_k^T\mathbf{g})$, where $\mathbf{g} = (\sum_{j\in S} d_j\mathbf{x}_j\mathbf{z}_j^T)^{-1}(\sum_{j\in U}\mathbf{x}_j - \sum_{j\in S} d_j\,\mathbf{x}_j)$ (although $d_k = 1$ in this application, we write the weighting equation more broady for future use). Unfortunately, there is no guarantee that the components of $\mathbf{g}$ (and thus some $\mathbf{z}_k^T\mathbf{g}$) will be nonnegative. That means that there is a possibility that some w_k will fall below unity, implying an element response probability greater than 1.

That never happens with any of our 5,000 *respondent subsets* based on the *100% population sets*. It does happen, however, with *respondent subsets* based on the smaller *20% population sets*. More details can be found in Section 5.

In our calibration-weight determinations, we use the following truncated form of the logistic back-link function:

$$f(\delta)= 1+ e^{-\delta} + \frac{1-e^{-2\delta}}{999+e^{-\delta}} = 1 + \frac{1+999e^{-\delta}}{999+e^{-\delta}} \qquad (19)$$

This restricts $f(\delta)$ to between 1000/999, which is slightly larger than 1, and 1000 (which is the restriction imposed on $1/p_k$ by equation (17)).

We conduct the following iterative search $\mathbf{g}$:

$$\mathbf{g}^{(r+1)} = \mathbf{g}^{(r)} + \{\sum_{k\in S} d_k f_1(\mathbf{z}_k{}^T\mathbf{g}^{(r)})\, \mathbf{x}_k\mathbf{z}_k{}^T\}^{-1} \{\sum_{k\in U} \mathbf{x}_k - \sum_{k\in S} d_k f(\mathbf{z}_k{}^T\mathbf{g}^{(r)})\mathbf{x}_k\}, \qquad (20)$$

where $f_1(\delta) = \partial f(\delta)/\partial\delta$. When no solution exists, usually because $\sum_{k\in S} d_k f_1(\mathbf{z}_k{}^T\mathbf{g}^{(r)})\mathbf{x}_k\mathbf{z}_k{}^T$ is not invertible, a model variable is dropped (which often corresponds to a zero or near zero in the diagonal of $\sum_{k\in S} d_k f_1(\mathbf{z}_k{}^T\mathbf{g}^{(r)})\mathbf{x}_k\mathbf{z}_k{}^T$) . We then replace $\mathbf{x}_k{}^T$ in equation (19) with $\hat{\mathbf{x}}_k{}^T = \mathbf{x}_k{}^T\hat{\boldsymbol{A}}$, where $\hat{\boldsymbol{A}}$, defined in equation (15b), can change with each iteration.

By truncating the back-link function in equation (17), we bound $f(\delta)^2 - f(\delta)$ away from zero and infinity. This allows $\sum_{j\in S}[\{d_j f(\mathbf{z}_j{}^T\mathbf{g}^{(r)})\}^2 - d_j f(\mathbf{z}_j{}^T\mathbf{g}^{(r)})]\mathbf{x}_j\mathbf{x}_j{}^T$ in equation (15b) to be invertible as long as $\mathbf{x}_j$ has full rank.

Notice, however, by constraining $f(\delta)$ to be above 1, we force the estimated response probability in a model group to be less than 1. This means that when the model and benchmark groups are (initially) the same, as is poststratification, using this back-link function would force the dropping of a component of $\mathbf{z}_k$ when there is 100% response among the respondents in a group.

5. The main results

In this section, we will call the model groups formed using true sales "response-guided response groups" and the calibration method based on them the "RGRG method." We will call calibration using the frame-defined benchmark groups as the response (i.e., model) groups "poststratification."

5.1. The 100% population sets

Although there are five 100% population sets, each have ten benchmark groups of population sizes 6001, 818, 685, 369, 52, 5507, 2116, 4658, 6728, and 1013. The smallest group is the one containing farms with the larest frame sales but no survey responses over the last five years.

The five population sets have response-guided-response-group population sizes ranging from 50 to 6603. The average response rate within such a model group ranges from 52 to 98%, and no group in any of the 5,000 respondent subsets is empty. In fact, we are aways able to compute calibration weights using the one step procedure: $w_k = d_k\{1 + \mathbf{z}_k^T(\sum_{j\in S} d_j\mathbf{x}_j\mathbf{z}_j^T)^{-1}(\sum_{j\in U}\mathbf{x}_j - \sum_{j\in S} d_j\mathbf{x}_j)\} = d_k\mathbf{z}_k^T(\sum_{j\in S} d_j\mathbf{x}_j\mathbf{z}_j^T)^{-1}\sum_{j\in U}\mathbf{x}_j = (\sum_{j\in U}\mathbf{x}_j^T)(\sum_{j\in S}\mathbf{z}_j\mathbf{x}_j^T)^{-1}d_k\mathbf{z}_k$ (since $\mathbf{z}_j^T\boldsymbol{\lambda} = 1$ for some $\boldsymbol{\lambda}$, $\sum_{j\in S} d_j\mathbf{x}_j = \sum_{j\in S} d_j\mathbf{x}_j\mathbf{z}_j^T\boldsymbol{\lambda}$; the rest follows almost immediately).

Scaling total sales to be 100, summary statistics averaged across 1000 respondent subsets are displayed in Table 1 for each population set and across the sets. They show that the absolute empirical bias from using the RGRG method is approximately 1/3 of that from poststratification. The empirical standard errors (SE) are slightly higher using the RGRG method rather than its rival, but the empirical root mean squared error (RMSE) from using RGRG is roughly 2/3 the size.

Since there is a definite bias when using incorrect models in the estimation, it is unclear whether equation (8) is supposed to estimate variance or mean squared error. The table suggests it consistently underestimates mean squared error while overestimating variance slightly. Using the "improved" version of the squared residuals suggested by equation (10) has very little effect on the standard-error/root-mean-squared-error estimates, tipping the rounding upward in some cases (not displayed).

5.2. The 20% population sets

The ten benchmark groups in the 20% population sets have population sizes ranging from 9 to 1310. The five generated populations have response-guided-response-group population sizes ranging from 7 to1314 and response rates again ranging from 52 to 98%. Although no response-guided response group ever contains no respondents, there are respondent subsets in which $\sum_{k\in S} d_k f_1(\mathbf{z}_k^T \mathbf{g}^{(r)})\, \mathbf{x}\mathbf{z}_k^T$ (for some iteration r), and on rare occasion, $\sum_{k\in S} d_k \mathbf{x}_k \mathbf{z}_k^T$, is not invertible. As a result, some component of $\mathbf{z}_k$ must be dropped.

One of the five populations, Population 2, produces all four respondent subsets where $\sum_{k\in S} d_k \mathbf{x}_k \mathbf{z}_k^T$ is not invertible and 99 others where $\sum_{k\in S} d_k f_1(\mathbf{z}_k^T \mathbf{g}^{(r)})\, \mathbf{x}_k \mathbf{z}_k^T$ is not. The other four populations produce 0, 1, 15, and 10 respondent subsets where $\sum_{k\in S} d_k f_1(\mathbf{z}_k^T \mathbf{g}^{(r)})\, \mathbf{x}_k \mathbf{z}_k^T$ is not invertible, respectively.

Perhaps the slightly larger empirical bias and empirical standard error for Population 2 under the RGRG method can be attributed to this pathology. Otherwise, the results for this set of populations is very similar to those for the 100% population sets. The empirical biases are about the same. The empirical standard errors are larger, as we would expect since the respondent subsamples have roughly 20% of their 100% counterparts, yet the model groups are the same for both (except when a component of $\mathbf{z}_k$ is dropped).

The estimated standard error/root-mean-squared errors are even closer to the empirical standard errors than before. Unfortunately, the "improved" variance estimates can not always be computed. One reason for this is that the denominator in equation (10) can be negative.

5.3 The respondent subsets for the 100% population under the actual response mechanism

Before trying to salvage the "improved" squared residuals in equation (10) under our simulations, we return to five 100% population sets and generate 1000 respondent subsets each using equation (17). We then calibrate using $f(.) = 1/p_k$, where

$$p_k = 0.001 + 0.998\{1 + \exp(\gamma_1 + \gamma_2\, logtvp2_k - \gamma_3\, s97)\}^{-1}.$$

Again scaling total sales to be 100, the overall empirical bias is 0.00257 and the overall empirical standard error is 0.613 (none of these results are displayed). The overall average estimate for the latter based on equation (8) is slightly smaller, 0.606. The same rough relationship between the estimated and empirical standard errors holds for each of the five populations. Using the improved squared residuals at most increases the estimated standard error by 0.001 (by s before rounding).

There is little reason to investigate the squared residuals in equation (10) further.

6. Discussion

The major result of the empirical investigation described in the last two section is that when response is a function of the variable of interest, using response-guided response groups rather than traditional frame-defined response groups can result in appreciable decreases in bias and root mean squared error even when neither the response nor prediction/measurement-error model fully justifying the methodology holds.

Using the frame-defined response groups generally results in a slightly smaller empirical standard error than its competitor, but this advantaged is overpowered by the increase in empirical bias. Similar findings (not displayed) result when the σ in equation (18) used to generate *true sales* from *frame sales* is increased from 1 to 10**.**

A noticeable amount of bias still remains relative to standard error for the 20% population sets even though sometimes the response-guided response groups are so small a model variable has to be dropped. Recall that a model variable is dropped either when $\sum_{k\in S} d_k \mathbf{x}_k \mathbf{z}_k^T$ is not of full rank and can not be inverted or when some $w_k = (\sum_{k\in U} \mathbf{x}_j^T)(\sum_{j\in S} d_j \mathbf{z}_j \mathbf{x}_j^T)^{-1} d_k \mathbf{z}_k$ is less than 1. The latter, which is more common, forces us to choose a form for the back-link function, $f(\delta)$ (*e.g.*, equation (19)). Otherwise, we can effectively set $f(\delta) = \delta$ and compute the w_k in one step using $w_k = (\sum_{k\in U} \mathbf{x}_j^T)(\sum_{j\in S} d_j \mathbf{z}_j \mathbf{x}_j^T)^{-1} d_k \mathbf{z}_k$.

As a practical matter, we suspect users will redefine their response-guided response groups (and corresponding benchmark groups) rather than be forced to select a nonlinear form for the back-link function and drop a model variable. Since the groups vary considerably in size in our analysis, there is some flexibility here to

form additional groups even for the smaller population size. Clearly, however, the number of response-guided model groups that can be used with a particular respondent data set will usually be smaller than the number of analogous frame-defined groups. More empirical work is needed on forming response-guided response groups.

We are suprised that the Kott-Brewer method for improving variance estimates has little effect. It may be that the distinction between an element's prediction/measurement-error model error ($\varepsilon_k = \tau_k - \boldsymbol{\xi}_k^T \boldsymbol{\Gamma}^{-1}\theta$) and its sample residual (e_k from equation (9)) plays a smaller role in variance/mean-squared-error estimation than other small-sample factors such as the randomness of the w_k due to their being functions of the random $\mathbf{x}_j$.

The asymptotic variance estimator in equation (8) always does a good job estimating empirical variance in our analyses. As an estimator for mean squared error, however, it works better the smaller the bias – better for the RGRG method than for poststratification.

Things will become more complicated when it is not a census that suffers nonresponse but a randomly selected sample from a population U. We noted in Section 3 that the prediction model in equation (11), $y_k = \mathbf{z}_k^T \theta + \tau_k$, and measurement-error model in equation (12), $\mathbf{x}_k = \mathbf{z}_k^T \boldsymbol{\Gamma} + \boldsymbol{\xi}_k$, combine to yield the standard prediction model in equation (3), $y_k = \mathbf{x}_k^T \beta + \varepsilon_k$.

It is often reasonable to assume $E\{\varepsilon_k | (\mathbf{x}_j; j \in U)\} = 0$ whether or not k is in the sample, even though this equality does not hold when conditioned on whether k responds. Thus, in the absence of nonresponse, it makes more sense to calibrate using the components of $\mathbf{x}_k$ as both the model and benchmark variables rather than having the component of $\mathbf{z}_k$, with their unknown population total, serve as the model variables. This suggests calibration, even when involving mutually exclusive groups, should be done in two steps, the first to calibrate the full sample to the population using only benchmark variables, and the second to calibrate the respondent sample to the full sample or population using response-guided model variables. We explore that suggestion below.

Using the *100% population set* described in the preceeding two sections, we create five populations using equation (18) with σ =1 to generate the total-sales values. We draw 25 simple random samples from each population and then 40

random subsamples of respondents from each original using the response probabilities in equation (17). We estimate total sales using a (pure) response-guided response-group method with benchmark and the response-guided response groups determined as in the previous sections. All the d_k are set equal to the population size divided by the original pre-nonresponse sample size. Label this value N/n.

We also estimate totals sales using only poststratification to the benchmark-group population sizes. Since a realized post-nonresponse sampling fraction, label it r_p/N_p for an element in group p, could equal or exceed the original sampling fraction, n/N, we use the simple poststratification weights, N_p/r_p, to compute these estimates rather than the approach described in earlier sections.

Finally, we estimate total sales by first poststratifying the original sample using the benchmark groups and then using the response-guided response groups to calibrate for nonresponse. After the first step, the calibration weight for every element in benchmark group p is N_p/n_p, where n_p is the original sample size in the benchmark group. The second step uses the response-guided response groups as described in the previous section with d_k set equal to N_p/n_p for elements in benchmark group p.

The results are summarized in Table 2. Notice that we often had to drop a model variable for $\sum_{k\in S} d_k \mathbf{x}_k \mathbf{z}_k^T$ to be invertible or for all $f(\mathbf{z}_k^T \mathbf{g})$ to be greater than unity. This happens nearly half the time when using the pure response-guided response-group method. Even the poststratified estimator could not always be calculated because one of the r_p was zero. Almost always, using the method involving two calibration results in the smallest empirical biases and root mean squared errors, although not by much.

A theoretical variance estimator is needed for such a two-stage calibration, especially with a more complicated calibration of the original pre-nonresponse sample. In this particular case, however, variance estimation can be done by setting $d_k = N_p/r_p = 1/\pi_k$ for respondents in benchmark-group p and then using equation (20) in Chang and Kott (2008).

REFERENCES

Bang H. and Robbins J.M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics, 61, 962-972.

Chang, T. and P. S. Kott (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 557-571. Available at http://www.nass.usda.gov/research/reports/cal_paper_rev3.pdf

Estevao, V. and Särndal, C.E. (2000). A functional form approach to calibration. Journal of Official Statistics, 16, 379-399.

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

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.

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

Kott, P. S. (2005). "No" Is the Easiest Answer: Using Calibration to Assess Nonignorable Nonresponse in the 2002 Census of Agriculture. Proceedings of the Section on Survey Research Methods, American Statistical Association,Washington DC. Another version available at http://www.nass.usda.gov/research/reports/JSM2005pap_alt_cover_3.pdf

Kott, P. S. and K. R. W. Brewer (2001). Estimating the Model Variance of a Randomization-consistent Regression Estimator. Proceedings of the Sectionon Survey Research Methods, American Statistical Association, Washington DC.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models.* Chapman and Hall, London.

Sautory, O. (2003) Calmar 2: A New Version of the Calmar Calibration Adjustment Program. Proceedings of the Statistics Canada Symposium 2003. http://www.statcan.ca/english/freepub/11-522-XIE/2003001/session13/sautory.pdf

Table 1. Comparing the two methods (Total sales scaled to equal 100)

		Response-guided reponse-group method				Poststratification method			
		Empirical bias	Empirical SE	Empirical RMSE	Estimated SE (RMSE?)	Empirical bias	Empirical SE	Empirical RMSE	Estimated SE (RMSE?)
	Population 1	- 0.25	0.46	0.52	0.47	- 0.74	0.44	0.86	0.44
	Population 2	- 0.23	0.51	0.56	0.52	- 0.73	0.49	0.88	0.49
100%	Population 3	- 0.26	0.53	0.60	0.54	- 0.73	0.52	0.89	0.52
Population	Population 4	- 0.25	0.39	0.46	0.40	- 0.72	0.37	0.81	0.38
Sets	Population 5	- 0.24	0.55	0.60	0.55	- 0.73	0.53	0.90	0.53
	Overall	- 0.25	0.49	0.55	0.50	- 0.73	0.47	0.87	0.48
	Population 1	- 0.26	0.66	0.71	0.65	- 0.79	0.63	1.01	0.61
	Population 2	- 0.27	0.73	0.78	0.71	- 0.81	0.68	1.06	0.65
20%	Population 3	- 0.27	0.65	0.70	0.66	- 0.79	0.61	1.09	0.61
Population	Population 4	- 0.21	0.71	0.75	0.72	- 0.82	0.65	1.04	0.63
Sets	Population 5	- 0.25	0.63	0.68	0.64	- 0.82	0.59	1.01	0.58
	Overall	- 0.25	0.68	0.72	0.68	- 0.80	0.63	1.02	0.62

Empirical Bias = $\left(\sum_{r=1}^{R} t_r - T\right) / R$, where t_r is the estimate for response-set r of R (= 1000 or 5000).

Empirical SE = $\sqrt{\left(\sum_{r=1}^{R} t_r - \left[\sum_{s=1}^{R} t_s / R\right]\right)^2 / (R-1)}$.

Empirical RMSE = $\sqrt{(Empirical\,Bias)^2 + (Empirical\,SE)^2}$.

Estimated SE = $\sqrt{\sum_{r=1}^{R} v_r / R}$, where v_r is computed using equation (8).

Table 2. Comparing three methods under simple random sampling

(Total sales scaled to equal 100)

	Percent of cases where a variable was dropped	Empirical bias	Empircal SE	Empirical RMSE
Pure response-guided reponse-group (RGRG) method				
Population 1	40.4	-0.25	2.01	2.02
Population 2	41.9	-0.38	2.02	2.05
Population 3	47.1	-0.28	2.14	2.16
Population 4	40.7	-0.54	1.72	1.80
Population 5	42.3	-0.45	1.64	1.70
Overall	42.5	-0.38	1.91	1.95
Poststratification method				
Population 1	*	-0.68	1.92	2.03
Population 2	*	-0.77	1.93	2.08
Population 3	*	-0.69	2.00	2.12
Population 4	*	-0.91	1.64	1.88
Population 5	*	-0.85	1.52	1.74
Overall	*	-0.78	1.81	1.97
Poststratification of the original sample followed by the RGRG method				
Population 1	3.8	-0.15	1.99	1.99
Population 2	3.4	-0.29	1.96	1.98
Population 3	9.7	-0.17	2.16	2.16
Population 4	5.0	-0.43	1.68	1.74
Population 5	5.3	-0.34	1.63	1.66
Overall	4.6	-0.27	1.89	1.92

* We excluded from these analyses the one case per population where a benchmark group had no responding elements in the sample.