Using Calibration to Fit Noncoverage and Nonresponse Models

Ted Chang and Phil Kott
University of Virginia
National Agricultural Statistical Service

Abstract

Calibration can be used to correct for sample nonresponse and frame undercoverage, as well as to assure that weighted estimates of the calibration variables match known or alternatively estimated population totals, called benchmarks. The quasi-randomization theory supporting this use treats response or coverage as an additional phase of random sampling (one that takes place before the sample is drawn in the case of undercoverage 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 calibration. Unfortunately, the variables in a reasonable quasi-random model are not necessarily the same as the calibration variables for which benchmark totals are available. Moreover, it is often prudent to keep the number of explanatory variables in a model small. We will address using calibration to adjust for nonreponse or undercoverage when then number of calibration variables exceeds the total number of explanatory model variables. Data from National Agricultural Statistical Service 2002 Census of Agriculture will be used to illustrate adjustment for nonresponse.

Keywords: Benchmarks; Quasi-random model; Model variables.

1. Introduction

Suppose $\mathbf{y}_i$ is a (column) p-vector of calibration variables for the i-th population element, and $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors, of length q and r, of model variables for a noncoverage model and nonresponse model, respectively. For reasons that will be made clear later, we assume that $\mathbf{x}_i$ and $\mathbf{z}_i$ have no components in common.

Suppose the probabilities of noncoverage and nonresponse are of the form $1-p_1(\mathbf{x}_i'\beta)$ and $1-p_2(\mathbf{z}_i'\gamma)$ respectively, for some vector parameters β and γ. If β and γ were known, the usual sample estimate of the vector of totals of the calibration variables would be

$$\widehat{t}_{\mathbf{y}}^{0} = \sum_{i\in\mathcal{R}} \frac{w_i}{p_1(\mathbf{x}_i'\beta)p_2(\mathbf{z}_i'\gamma)}\mathbf{y}_i, \qquad (1)$$

where w_i is the sampling weight and $\mathcal{R}$ is the set of respondents.

If $T_{\mathbf{y}}$ is a vector of calibration target values consisting of known, or previously estimated, population totals, then β and γ could be estimated from the data using the calibration equations

$$T_{\mathbf{y}} = \sum_{i\in\mathcal{R}} \frac{w_i}{p_1(\mathbf{x}_i'\widehat{\beta})p_2(\mathbf{z}_i'\widehat{\gamma})}\mathbf{y}_i. \qquad (2)$$

If the number p of calibration variables equals the number $q+r$ of model variables, equations (2) will usually be sufficient to determine $\widehat{\beta}$ and $\widehat{\gamma}$. On the other hand, if $p<q+r$, $\widehat{\beta}$ and $\widehat{\gamma}$ will be underdetermined by (2).

If β and γ were known, it is unlikely that $\widehat{t}_{\mathbf{y}}^{0}$ of (1) would equal $T_{\mathbf{y}}$ exactly due to sampling variability. The vectors $\widehat{t}_{\mathbf{y}}^{0}$ and $T_{\mathbf{y}}$ should be close, however. With this in mind, we suggest that (1), and its child (2), be viewed as nonlinear regression-type equations

$$T_{\mathbf{y}} = \widehat{t}_{\mathbf{y}}(\beta,\gamma) + \epsilon \qquad (3)$$

$$\widehat{t}_{\mathbf{y}}(\beta,\gamma) = \sum_{i\in\mathcal{R}} \frac{w_i}{p_1(\mathbf{x}_i'\beta)p_2(\mathbf{z}_i'\gamma)}\mathbf{y}_i$$

where ϵ is a p-vector of "errors". In this setup, it is desirable that $p>q+r$ and, indeed, the more calibration targets the merrier. The parameters β and γ can be estimated by minimizing an objective function of the form

$$\rho(\beta,\gamma) = (T_{\mathbf{y}} - \widehat{t}_{\mathbf{y}}(\beta,\gamma))'\mathbf{W}(T_{\mathbf{y}} - \widehat{t}_{\mathbf{y}}(\beta,\gamma)) \qquad (4)$$

for some appropriately chosen $p\times p$ positive definite matrix $\mathbf{W}$.

Under mild conditions, minimizing the objective function will yield consistent estimators for the parameters no matter what one chooses for $\mathbf{W}$. Nevertheless, 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-covariance matrix $\mathbf{V}$ of ϵ. When this choice of $\mathbf{W}$ depends on knowing (β,γ), we propose an iterative procedure. Given a guess $\widehat{\theta}_0$ of (β,γ), we can linearize the regression (3) at $\widehat{\theta}_0$. The solution to the linearized regression is the next guess $\widehat{\theta}_1$. This

procedure is described more thoroughly in Section 2. It can be shown to be optimal under a criterion given by Thompson (1997).

An obvious candidate for $\mathbf{V}$ is $\widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma))$, a sample estimate of the design-based variance of $\widehat{t}_{\mathbf{y}}(\beta,\gamma)$. If the sampling scheme is without replacement so that $\widehat{t}_{\mathbf{y}}(\beta,\gamma)$ of (3) is the Horvitz Thompson estimator, then $\widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma))$ is given by

$$\widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma)) = \sum_{i\in\mathcal{R}} \frac{1-p_i}{p_i^2\pi_{i|\mathcal{C}}}\mathbf{y}_i\mathbf{y}_i^{'} \qquad (5)$$

$$+ \sum_{i,j\in\mathcal{R}} \frac{\pi_{ij|\mathcal{C}} - \pi_{i|\mathcal{C}}\pi_{j|\mathcal{C}}}{\pi_{ij|\mathcal{C}}\pi_{i|\mathcal{C}}\pi_{j|\mathcal{C}}p_ip_j}\mathbf{y}_i\mathbf{y}_j^{'},$$

where $p_i = p_1(\mathbf{x}_i^{'}\beta)p_2(\mathbf{z}_i^{'}\gamma)$

In (5) we envision that the respondents $\mathcal{R}$ are a three phase sample from the target population $\mathcal{U}$. In the first phase, the coverage population $\mathcal{C}$ is a Poisson sample (see Särndal *etal* (1992), page 85) from $\mathcal{U}$ with sampling probabilities $p_1(\mathbf{x}_i^{'}\beta)$. In the second phase, the sample $\mathcal{S}$ is a sample without replacement from $\mathcal{C}$ with inclusion probabilities $\pi_{i|\mathcal{C}} = Pr[\ i \in \mathcal{S} \mid \mathcal{C}]$ and $\pi_{ij|\mathcal{C}} = Pr[\ i,j \in \mathcal{S} \mid \mathcal{C}]$. These probabilities normally depend upon $\mathcal{C}$ and the notation is meant to so indicate. Furthermore, in this case, $w_i = \pi_{i|\mathcal{C}}^{-1}$. Finally we assume that $\mathcal{R}$ is a Poisson sample from $\mathcal{S}$ with sampling probabilities $p_2(\mathbf{z}_i^{'}\gamma)$.

Section 3 gives the derivation of (5) as well as the formulas of $\widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma))$ for other designs of interest.

If the targets $T_{\mathbf{y}}$ are previously estimated population totals, then it would be appropriate to set

$$\mathbf{V} = \widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma)) + Var(T_{\mathbf{y}})$$

where $Var(T_{\mathbf{y}})$ is some estimate of the errors in $T_{\mathbf{y}}$. Notice that $Var(T_{\mathbf{y}})$ is not necessarily diagonal, so that the components of the target vector $T_{\mathbf{y}}$ can be, and usually are, correlated. In the case of the National Agricultural Statistical Service surveys, the targets are often derived from expert opinions, which means that $Var(T_{\mathbf{y}})$ must be determined heuristically. The literature on the choice of informative Bayesian priors gives many suggestions which can be adapted for this purpose.

In Section 4, we give an estimator for the quasi-design-based errors of calibrated estimates of the totals of a (vector) variable of interest $\mathbf{u}$. This variance estimate will include the variability due to the sampling design, the variability under the Poisson undercoverage/nonresponse model, and the variability in the estimates of β and γ.

Finally, in Section 5, we will give the results of an experiment to calibrate the 2002 Census of Agriculture for nonresponse.

2. Some Details

Recall that we assume that the probabilities of noncoverage and nonresponse are of the form $1-p_1(\mathbf{x}^{'}\beta)$ and $1-p_2(\mathbf{z}^{'}\gamma)$ respectively for some known functions p_1 and p_2 and unknown β and γ. We also assume that the components of $\mathbf{x}$ in the nonresponse model and those of $\mathbf{z}$ in the noncoverage model are distinct. Indeed, the pair $(\mathbf{x},\mathbf{z})$ should not be close to collinear. We make this second assumption because otherwise

$$\begin{aligned} p_1(\mathbf{x}^{'}\beta)p_2(\mathbf{z}^{'}\gamma) &\approx p_1(\mathbf{x}^{'}\beta_0)p_2(\mathbf{z}^{'}\gamma_0) \\ &+ p_1(\mathbf{x}^{'}\beta_0)p_2^{'}(\mathbf{z}^{'}\gamma_0)\mathbf{z}^{'}(\gamma-\gamma_0) \\ &+ p_1^{'}(\mathbf{x}^{'}\beta_0)p_2(\mathbf{z}^{'}\gamma_0)\mathbf{x}^{'}(\beta-\beta_0) \end{aligned}$$

As a result, (β,γ) 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 $(\widehat{\beta},\widehat{\gamma})$.

Write $\theta = [\beta^{'} \quad \gamma^{'}]^{'}$. Given a guess $\widehat{\theta}_0$ of θ and matrix $\mathbf{W}(\widehat{\theta}_0)$ we linearize (3) at $\widehat{\theta}_0$ and obtain

$$T_{\mathbf{y}} - \widehat{t}_{\mathbf{y}}(\widehat{\theta}_0) = \widehat{\mathbf{H}}(\widehat{\theta}_0)\left(\theta - \widehat{\theta}_0\right) + \epsilon \qquad (6)$$

where $\widehat{\mathbf{H}}(\widehat{\theta}_0)$ is the $p \times (q+r)$ matrix

$$\widehat{\mathbf{H}}(\widehat{\theta}_0) = \frac{\partial \widehat{t}_{\mathbf{y}}(\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_0}. \qquad (7)$$

The (weighted) linear regression estimate $\widehat{\theta}_1$ corresponding to (6) minimizes the objective function $\mathbf{U}^{'}\mathbf{W}(\widehat{\theta}_0)\mathbf{U}$ where $\mathbf{U} = T_{\mathbf{y}} - \widehat{t}_{\mathbf{y}}(\widehat{\theta}_0) - \widehat{\mathbf{H}}(\widehat{\theta}_0)(\theta - \widehat{\theta}_0)$. It is given by

$$\widehat{\theta}_1 - \widehat{\theta}_0 = \left[\widehat{\mathbf{H}}^{'}\mathbf{W}\widehat{\mathbf{H}}\right]^{-1}\left[\widehat{\mathbf{H}}^{'}\mathbf{W}\left(T_{\mathbf{y}} - \widehat{t}_{\mathbf{y}}(\widehat{\theta}_0)\right)\right] \qquad (8)$$

where the matrices $\mathbf{H}$ and $\mathbf{W}$ are evaluated at $\widehat{\theta}_0$. $\widehat{\theta}_1$ is the next guess.

If $\mathbf{W}$ is an estimate $\widehat{Var}(\widehat{t}_{\mathbf{y}}(\widehat{\theta}))$ of the covariance matrix of $\widehat{t}_{\mathbf{y}}(\widehat{\theta})$, an alternative derivation and justification of the update equation (8) can be provided using the ideas of Thompson (1997). We consider the equations $\widehat{t}_{\mathbf{y}}(\theta) - T_{\mathbf{y}} = 0$ as p estimating equations for the $q+r$ coefficients θ. If $\mathbf{A}$ is a $(q+r)\times p$ matrix of constants, let $\widehat{\theta}_A$ denote the solution to the estimating equations

$$\mathbf{A}\widehat{t}_{\mathbf{y}}(\theta) = \mathbf{A}T_{\mathbf{y}}. \qquad (9)$$

We seek the matrix $\mathbf{A}^*$ such that $\widehat{\theta}_{A^*}$ is optimal in asymptotic variance. Numerical solution for $\mathbf{A}^*$ and $\widehat{\theta}_{A^*}$ leads to the update equation (8).

3. Some Design Based Covariance Matrices

We start with the proof of equation (5)

Proposition. Suppose that the coverage population $\mathcal{C}$ is a Poisson sample from population $\mathcal{U}$ with sampling probabilities $p_{1i} = p_1(\mathbf{x}_i'\beta)$, that the sample $\mathcal{S}$ is a sample without replacement from $\mathcal{C}$ with inclusion probabilities $\pi_{i|\mathcal{C}} = Pr[\ i \in \mathcal{S} \mid \mathcal{C}]$ and $\pi_{ij|\mathcal{C}} = Pr[\ i,j \in \mathcal{S} \mid \mathcal{C}]$, and finally that the responding sample $\mathcal{R}$ is a Poisson sample from $\mathcal{S}$ with sampling probabilities $p_{2i} = p_2(\mathbf{z}_i'\gamma)$. Let $w_i = \pi_{i|\mathcal{C}}^{-1}$. Then

$$\widehat{t}_{\mathbf{y}}(\beta,\gamma) \quad = \quad \sum_{i\in\mathcal{R}} \frac{w_i}{p_1(\mathbf{x}_i'\beta)p_2(\mathbf{z}_i'\gamma)}\mathbf{y}_i$$

is an unbiased estimate of $T_{\mathbf{y}} = \sum_{i\in\mathcal{U}} \mathbf{y}_i$ and an unbiased estimate of its variance is given by equation (5).

Proof: Suppose temporarily that there is no noncoverage and hence $\mathcal{C} = \mathcal{U}$ and we write π_i and π_{ij} in place of $\pi_{i|\mathcal{C}}$ and $\pi_{ij|\mathcal{C}}$. Then Särndal *etal* (1992) equation (9.3.7) yields

$$\widehat{Var}_{db}(\widehat{t}_{\mathbf{y}}(\gamma)) = \sum_{i\neq j\in\mathcal{R}} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j p_{2i}p_{2j}}\mathbf{y}_i\mathbf{y}_j' \qquad (10)$$

$$+\sum_{i\in\mathcal{R}} \frac{1-p_{2i}}{p_{2i}^2\pi_i^2}\mathbf{y}_i\mathbf{y}_i' + \sum_{i\in\mathcal{R}} \frac{1-\pi_i}{\pi_i^2 p_{2i}}\mathbf{y}_i\mathbf{y}_i'$$

$$= \sum_{i\in\mathcal{R}} \frac{1-p_{2i}}{p_{2i}^2\pi_i}\mathbf{y}_i\mathbf{y}_i' + \sum_{i,j\in\mathcal{R}} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j p_{2i}p_{2j}}\mathbf{y}_i\mathbf{y}_j'$$

Now returning to the general case and reviving the notation $\pi_{i|\mathcal{C}}$ and $\pi_{ij|\mathcal{C}}$, let $\widehat{t}_C = \sum_{i\in\mathcal{C}} \frac{\mathbf{y}_i}{p_1(\mathbf{x}_i'\beta)}$. Then

$$E\left(\widehat{t}_{\mathbf{y}}(\beta,\gamma)|\mathcal{C}\right) = \widehat{t}_C$$

$$Var\left(\widehat{t}_{\mathbf{y}}(\beta,\gamma)\right) = Var\left(\widehat{t}_C\right) + E\left(Var\left(\widehat{t}_{\mathbf{y}}(\beta,\gamma)|\mathcal{C}\right)\right)$$

$$Var\left(\widehat{t}_C\right) \quad = \quad \sum_{i\in\mathcal{U}} \frac{1-p_{1i}}{p_{1i}}\mathbf{y}_i\mathbf{y}_i'$$

$$= \quad E\left(\sum_{i\in\mathcal{R}} \frac{1-p_{1i}}{p_{1i}^2 p_{2i}\pi_{i|\mathcal{C}}}\mathbf{y}_i\mathbf{y}_i'\right)$$

$$Var\left(\widehat{t}_{\mathbf{y}}(\beta,\gamma)|\mathcal{C}\right) = \sum_{i\in\mathcal{R}} \frac{1-p_{2i}}{p_{2i}^2\pi_{i|\mathcal{C}}}\frac{\mathbf{y}_i}{p_{1i}}\frac{\mathbf{y}_i'}{p_{1i}}$$

$$+\sum_{i,j\in\mathcal{R}} \frac{\pi_{ij|\mathcal{C}}-\pi_{i|\mathcal{C}}\pi_{j|\mathcal{C}}}{\pi_{ij|\mathcal{C}}\pi_{i|\mathcal{C}}\pi_{j|\mathcal{C}}p_{2i}p_{2j}}\frac{\mathbf{y}_i}{p_{1i}}\frac{\mathbf{y}_j'}{p_{1j}},$$

where the last equation follows by applying (10) to the variables $p_{1i}^{-1}\mathbf{y}_i$. Thus (5) follows. □

As a second example we note:

Proposition. Considered a stratified multistage design with PSU's chosen with replacement. Let $h = 1,\ldots,H$ index the strata, and for each h, let $\mathcal{U}_{Ih}$ denote the population of PSU's in stratum h. For $i \in \mathcal{U}_{Ih}$, let $\mathcal{U}_{hi}$ denote the elements of $\mathcal{U}$ in PSU (h,i) and let the sampling weights be w_{hij} for $j \in \mathcal{U}_{hi}$.

Suppose in stratum h, n_h PSU's are chosen with replacement and probability (actually expected count) $n_h z_{hi}$ and let $\mathcal{R}_{hi} = \mathcal{U}_{hi} \cap \mathcal{R}$. Let

$$\widehat{t}_{\mathbf{y}hi}(\beta,\gamma) = n_h z_{hi} \sum_{j\in\mathcal{R}_{hi}} \frac{w_{hij}}{p_1(\mathbf{x}_{hij}'\beta)p_2(\mathbf{z}_{hij}'\gamma)}\mathbf{y}_{hij}$$

$$\widehat{t}_{\mathbf{y}h}(\beta,\gamma) = n_h^{-1}\sum_{i=1}^{n_h} z_{hi}^{-1}\widehat{t}_{\mathbf{y}hi}(\beta,\gamma)$$

Then $\widehat{t}_{\mathbf{y}}(\beta,\gamma) = \sum_h \widehat{t}_{\mathbf{y}h}(\beta,\gamma)$ is an unbiased estimate of $T_{\mathbf{y}}$ and its variance can be unbiasedly estimated by $\sum_h n_h^{-1}s_h^2$ where

$$s_h^2 = (n_h-1)^{-1}\sum_{i=1}^{n_h}(z_{hi}^{-1}\widehat{t}_{\mathbf{y}hi}(\beta,\gamma) - \widehat{t}_{\mathbf{y}h}(\beta,\gamma))^2.$$

Proof: Let $T_{\mathbf{y}hi} = \sum_{j\in\mathcal{U}_{hi}} \mathbf{y}_{hij}$. Then $E[\widehat{t}_{\mathbf{y}hi}(\beta,\gamma) \mid (h,i) \in \text{sample}] = T_{\mathbf{y}hi}$. The result follows from Särndal *etal* (1992), page 151. □

Finally, notice that the estimates $\widehat{\beta}$ and $\widehat{\gamma}$ are unchanged if $Var_{db}(\widehat{t}_{\mathbf{y}}(\beta,\gamma))$ is only estimated up to a multiplicative constant. This suggests invoking the spirit of design effects and using the variance from simple random sampling:

$$\overline{\mathbf{y}} = n^{-1}\sum_{i\in\mathcal{R}} p_i^{-1}\mathbf{y}_i$$

$$\widehat{\mathbf{V}} = \frac{N^2}{n(n-1)}\sum_{i\in\mathcal{R}}(p_i^{-1}\mathbf{y}_i - \overline{\mathbf{y}})'(p_i^{-1}\mathbf{y}_i - \overline{\mathbf{y}})$$

where, as before, $p_i = p_1(\mathbf{x}_i'\beta)p_2(\mathbf{z}_i'\gamma)$. In practice, the scalar multiple, $\frac{N^2}{n(n-1)}$, can be dropped from $\widehat{\mathbf{V}}$.

4. Variance of Calibrated Estimates of Population Totals

Let $\mathbf{u}_i$ be a vector of variables of interest. Our calibrated estimate for the total $t_{\mathbf{u}}$ is $\widehat{t}_{\mathbf{u}}(\widehat{\beta},\widehat{\gamma})$ where

$$\widehat{t}_{\mathbf{u}}(\beta,\gamma) = \sum_{i\in\mathcal{R}} \frac{w_i}{p_i(\beta,\gamma)}\mathbf{u}_i$$

with $p_i(\beta,\gamma) = p_1(\mathbf{x}_i'\beta)p_2(\mathbf{z}_i'\gamma)$. Let

$$\widehat{\mathbf{H}}_\mathbf{u} = \frac{\partial \widehat{t}_\mathbf{u}(\beta,\gamma)}{\partial(\beta,\gamma)}(\widehat{\beta},\widehat{\gamma})$$

$$\mathbf{b} = \widehat{\mathbf{H}}_\mathbf{u}\left[\widehat{\mathbf{H}}'\mathbf{V}^{-1}\widehat{\mathbf{H}}\right]^{-1}\mathbf{A}^*,$$

where $\widehat{\mathbf{H}}$ is the matrix of partial derivatives (7) and $\mathbf{A}^* = \mathbf{H}'\mathbf{V}^{-1}$. Here $\widehat{\mathbf{H}}_\mathbf{u}$, $\widehat{\mathbf{H}}$, and $\mathbf{A}^*$ are evaluated at $(\widehat{\beta},\widehat{\gamma})$.

Then to first approximation

$$\widehat{t}_\mathbf{u}(\widehat{\beta},\widehat{\gamma}) \approx \mathbf{b}\widehat{t}_\mathbf{y}(\widehat{\beta},\widehat{\gamma}) + \sum_{i\in\mathcal{R}}\frac{w_i}{p_i(\beta,\gamma)}(\mathbf{u}_i - \mathbf{b}\mathbf{y}_i)$$

$$\approx \mathbf{b}t_\mathbf{y} + \sum_{i\in\mathcal{R}}\frac{w_i}{p_i(\beta,\gamma)}(\mathbf{u}_i - \mathbf{b}\mathbf{y}_i).$$

This suggests that the design-based errors in $\widehat{t}_\mathbf{u}(\widehat{\beta},\widehat{\gamma})$ can be calculated using the earlier formulas for $\widehat{Var}_{db}$ by substituting $\mathbf{u}_i - \mathbf{b}\mathbf{y}_i$ for $\mathbf{y}_k$ and estimating all p_i using $(\widehat{\beta},\widehat{\gamma})$.

5. Example

We consider here the calibration of the 2002 Census of Agriculture for nonresponse. Ignoring the incompleteness of the list from which that agricultural census was enumerated, we apply equation (5) with all $\pi_{i|\mathcal{C}}$, $\pi_{ij|\mathcal{C}}$, and coverage probabilities p_1 set equal to 1. This is what the National Agricultural Statistics Service (NASS) did when adjusting for unit nonresponse.

Before enumeration, each farm in the frame was assigned a value for its expected annual sales. Together with whether or not the farm responded to a survey since the 1997 Census of Agriculture, NASS used expected annual sales to divide the farm-frame population into five response groups. These groups were farms with expected annual sales less than \$2,500, expected annual sales between \$2,500 and \$10,000, expected annual sales between \$10,000 and \$50,000 and a survey response since 1997, expected annual sales over \$50,000 and a survey response since 1997, expected annual sales over \$10,000 and no survey response since 1997.

Table 1: Fitted Response Model Coefficients

	CA	*DE*
z_1 *intercept*	3.748	2.316
z_2 *log sales*	-0.2341	-0.09644
z_3 *response 97*	0.3841	0.1543

We use indicator variables for these five response groups as our calibration variables. NASS used them as both calibration and response-model variables. See Kott (2005) for more details.

In contrast to NASS's approach, we model response using three $\mathbf{z}$-variables: z_1 an intercept, z_2 the logarithm of the actual annual sales in 2002, truncated to the range \$1000 to \$100,000, and z_3 an indicator variable for whether or not the farm responded to a survey since the 1997 Census of Agriculture. Whereas the *expected* annual sales was calculated by NASS, and hence known for all farms, *actual* annual sales is calculated from 2002 Census responses and hence known only for respondents. A logistic link was used. Thus

$$p_2(\mathbf{z}'\gamma) = (1+\exp(-\eta))^{-1}$$

$$\eta = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2.$$

Two states were considered: California and Delaware. These two states were chosen to be as dissimilar as possible. The fitted response model coefficients $\widehat{\gamma}$ are given in Table 1.

The calibration targets $T_\mathbf{y}$ and their fitted values $\widehat{t}_\mathbf{y}(\widehat{\gamma})$ are given in Table 2.

We use calibration to estimate the total number of active farms. Hence our u variable is a 0-1 variable for being an active farm. We compare our technique here with postratification using the same five strata as the response groups above. The results are given in Table 3, with the standard errors, calculated as in Section 4, in parentheses.

The calibration standard errors are slightly larger than those given by poststratification. Notice, however, that the poststratification nonresponse weights necessarily depend upon the NASS assigned *expected sales* y and not upon the actual sales z. Furthermore, poststratification weights will change abruptly as the expected sales goes through the boundaries \$2500, \$10,000, and \$50,000 and the poststratification standard errors are calculated assuming this somewhat unrealistic model. Thus the poststratification standard errors are probably too low. Calibration allows one to fit a more realistic nonresponse

Table 2: Calibration Targets $T_\mathbf{y}$ and Corresponding Fitted Values $\widehat{t}_\mathbf{y}(\widehat{\gamma})$

	y_1	y_2	y_3	y_4	y_5
CA $T_\mathbf{y}$	21804	14622	14309	14777	4769
CA $\widehat{t}_\mathbf{y}(\widehat{\gamma})$	21861	14578	14274	14816	4752
DE $T_\mathbf{y}$	628	369	334	517	216
DE $\widehat{t}_\mathbf{y}(\widehat{\gamma})$	638.9	370.3	311.5	535.4	207.9

Table 3: Comparison of Estimates, with Standard Errors, of Total Number of Farms. Estimation by Poststratification Using Expected Sales, Calibration (This Paper) Using Actual Sales, and Calibration (as in Kott (2005)) Using Actual Sales as a Categorical Vairable.

	Poststratification		*Calibration (This Paper)*		*Calibration (Kott (2005))*	
	$\widehat{t}_{\mathbf{u}}(\widehat{\gamma})$	*s.e.*	$\widehat{t}_{\mathbf{u}}(\widehat{\gamma})$	*s.e.*	$\widehat{t}_{\mathbf{u}}(\widehat{\gamma})$	*s.e.*
CA	45312.5	45.8	46178.8	56.3	46181.3	165.6
DE	1390.9	9.2	1400.6	11.0	1416.1	18.7

model with a very reasonable increase in standard error.

We also give the breakdown of the example in Kott (2005) to the states CA and DE. In that example the model variables were 5 indicator variables corresponding to 5 classes defined analogously to the 5 calibration variable classes with actual sales replacing expected sales in the definition of the 5 model variable classes. Since, for this example, the model variables are all mutually exclusive indicator variables, all link functions are equivalent. In this example, the standard errors are substantially higher.

6. Concluding Remarks

A companion paper, Kott (2005), discusses calibration for nonresponse when there are exactly the same number of calibration and nonresponse model variables. As Kott (2004) points out, a very similar approach can be used to treat calibration for noncoverage (or overcoverage with slight modification).

This paper provides extends Kott (2005) in allowing the number of model variables to be less, but not more, than the number of calibration variables. In effect, these two papers allow for a complete separation of the calibration and model variables. Thus calibration to correct for nonresponse and/or noncoverage is made more realistic through more realistic modeling of the nonresponse/noncoverage weights.

In practice, calibrations for nonresponse and noncoverage are often done separately. This is because nonresponse calibration uses benchmarks from an incomplete list frame, while noncoverage calibration uses benchmarks wholly or partially determined from outside sources.

An example of where nonresponse and noncoverage calibration has been done simultaneously can be found in Crouse and Kott (2004). Rather than assuming the components of the response and coverage models were distinct, Crouse and Kott supposed that the components were, or could be, exactly the same. They were able to do this by assuming both models had the form $p(.) = exp(.)$ so that $p_i = p_1(\mathbf{x}_i^{'}\beta)p_2(\mathbf{z}_i^{'}\gamma)$ was also of that form. Although in their setup, the effects of nonresponse and noncoverage can not separated, the focus of Crouse and Kott (2004) was the primary goal of survey sampling - the estimation of totals (or functions of totals) for survey variables.

Much work is needed in determining how to select model and calibration variables in practice, especially when the benchmark targets are themselves potentially subject to sampling and measurement errors.

References

Crouse, C and Kott, P. (2004), "Evaluating alternative calibration schemes for an economic survey with large nonresponse," *Proceedings of the ASA Survey Reseach Methods Section.*

Kott, P. (2004), "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. (2005), "'No' is the easiest answer: using calibration to assess nonignorable nonresponse in the 2002 census of agriculture," *Proceedings of the ASA Survey Reseach Methods Section* (this volume).

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

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