AN OVERVIEW OF CALIBRATION WEIGHTING

Phillip S. Kott
National Agricultural Statistics Service

KEY WORDS: Unbiased, Randomization consistent, Raking, Nonresponse, Jackknife Variance Estimator, Coverage Adjustment.

I. Introduction

Suppose one wanted to estimate totals for a number of target variables based on data from a probability sample. If we knew the selection probability, π_k, for each sample element k in the sample S, then we could estimate any population total, $T_y = \sum_U y_k$, where U denotes the population, with the expansion estimator $t_{y\_E} = \sum_S y_k/\pi_k = \sum_U y_k I_k/\pi_k$, where $I_k = 1$ when $k \in S$ and 0 otherwise. Treating the I_k as random variables, it is easy to see that t_{y_E} is an unbiased estimator for T_y. We call properties arising when the I_k are treated as random variables *randomization-based*.

We can also write $t_{y\_E} = \sum_U a_k y_k = \sum_S a_k y_k$, where $a_k = I_k/\pi_k$ is called the sampling weight of element k. This same formula applies for any variable y_k about which we can collect data whenever k is in the sample.

Deville and Särndal (1992) coined the term "calibration estimator" to describe an estimator of the form $t_{y\_CAL} = \sum_S w_k y_k$, where $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_\mathbf{x}$ for some row vector of auxiliary variables, $\mathbf{x}_k = (x_{1k}, ..., x_{pk})$, about which $T_\mathbf{x}$ is known. Since there is generally a continuum of sets $\{w_k \mid k \in S\}$ that satisfy the *calibration equation*:

$$\sum_{k\in S} w_k \mathbf{x}_k = T_\mathbf{x}, \qquad (1)$$

Deville and Särndal required that the difference between $\{w_k \mid k \in S\}$ and $\{a_k \mid k \in S\}$ minimize some loss function.

The univariate components of equation (1):

$$\sum_{k\in S} w_k x_{pk} = \sum_{k\in U} x_{pk} \text{ for } p = 1, ..., P,$$

are sometimes called the "calibration equations." Chambers (1996) referred to a set of weights that satisfies the calibration equation(s), whether or not they minimize a loss function, as "case-based."

As with the expansion estimator, the same set of calibration weights can be used no matter what the variable of interest, y_k. When the particular y_k is a linear combination of the components of $\mathbf{x}_k$ for all $k \in U$, say $\mathbf{x}_k\beta$, then t_{y_CAL} equals T_y exactly. This is a great strength of calibration weighting and the reason behind why the calibration estimator is often much more efficient (has a smaller mean squared error) than the expansion estimator.

Another strength of calibration weighting is that $\{w_k \mid k \in S\}$ and $\{a_k \mid k \in S\}$ must be close since their difference is in some sense minimized. As a result, with a sufficiently large sample, t_{y_CAL} is close to randomization unbiased no matter what the y-variable is as long as it obeys reasonable regularity conditions to be described in the next section.

Since t_{y_CAL} estimates T_y perfectly when $y_k = \mathbf{x}_k\beta$ exactly, it is reasonable to expect t_{y_CAL} to be a good estimator when y_k and $\mathbf{x}_k\beta$ are close. This can be formalized by assuming the y_k are random variables satisfying the linear *prediction* model:

$$y_k = \mathbf{x}_k\beta + \epsilon_k, \qquad (2)$$

where $E(\epsilon_k \mid \{\mathbf{x}_g \mid g \in S\}, \{I_g \mid g \in U\}) = 0$ for all $k \in U$. Under this model, it is easy to see that t_{y_CAL} is an unbiased estimator for T_y in the sense that $E_\epsilon(t_{y\_CAL} - T_y) = 0$ (suppressing the conditioning for notational convenience); the subscript ϵ refers to treating the ϵ_k as random variables (and the I_k as fixed constants).

One problem with *model-based* analysis in practice is that we are usually interested in estimating totals for variety of target variables at the same time. It is often unreasonable to assume that different variables satisfy the same linear model.

This problem can be made to all but disappear. Suppose we had postulated separate models for J different target variables, $y_{1k}, ..., y_{Jk}$:

$$y_{jk} = \mathbf{x}_{jk}\beta_j + \epsilon_{jk},$$

where $\mathbf{x}_{jk}$ is a p_j-component row vector, and $E(\epsilon_{jk} \mid \{\mathbf{x}_{1g}, ..., \mathbf{x}_{Jg} \mid g \in S\}, \{I_g \mid g \in U\}) = 0$ for all $k \in U$. It is obvious that the model in equation (2) still holds with $\mathbf{x}_k$ now equal to $(\mathbf{x}_{1k}, ..., \mathbf{x}_{Jk})$. Duplicated and singular components of $\mathbf{x}_k$ can be pruned with no practical effect on the model (a singular component is a linear combination of other components).

A simple example is the following. Suppose y_{1k} is the current planted corn acres for farm k, and y_{2k} the farm's current planted wheat acres. Several years ago, all the farms in the population provided their annual corn and wheat acres to the Census of Agriculture. Denoting these previous values for farm k as x_{1k} and x_{2k}, respectively, the combined linear model inherent in

calibration takes the form:

$$y_{jk} = (1 \;\; x_{1k} \;\; x_{2k})(\beta_{0j} \;\; \beta_{1j} \;\; \beta_{2j})' + \epsilon_{jk}$$

for j = 1 or 2. Notice that the $\mathbf{x}_k$-vector is common to both the model for corn and wheat. The β-vector is not. The common $\mathbf{x}_k$-vector allows the creation of a common set of calibration weights for each target variable.

Calibration has its drawbacks. In the simple farm example, it may be reasonable to assume that β_{12} and β_{21} are zero, that corn in the census year has no effect on the current amount of planted wheat, and that census-year wheat has no effect on current-year corn. By explicitly assuming these equalities in estimation, efficiency is likely to increase. Unfortunately, calibration does not allow us to do that. It is the price we pay for developing a single set of weights for all target variables.

Although they coined the term, Deville and Särndal were not the first to note that a single set of weights, $\{w_k \mid k \in S\}$, could be constructed so that the resulting estimator, t_{y_CAL}, is

1) model unbiased under equation (2) as long as the $\mathbf{x}$-vector has the same set of components for every target variable of interest. and,
2) nearly randomization unbiased.

Huang and Fuller (1976) developed software to produce what are now called "calibration weights." Their approach usually returns the randomization-consistent regression estimator with the added constraint that each w_k be bounded by $(1 - M)a_k \le w_k \le (1 + M)a_k$ for a specified M.

Poststratification is a form of calibration that preceded Huang and Fuller by decades. It is most often used to adjust for unit nonresponse in the sample or coverage errors in the sampling frame, but in the discussion below we assume a perfect frame and complete response. We will return to postratification as a method for handling unit nonresponse briefly in Section 8.

Suppose the components of $\mathbf{x}_k$ are binary classification variables such that $x_{pk} = 1$ when k is in Class p and 0 otherwise. In a human population, for example, we can have $x_{1k} = 1$ and $x_{2k} = 0$ when k is male, and $x_{1k} = 0$ and $x_{2k} = 1$ when k is female. When each k is in one and only one of the P classes, as in the example, a poststratified estimator performs a simple ratio adjustment, setting each $w_k = (N_p / \sum_S a_j x_{pj})a_k$, when k is in both the sample and in class p, and N_p is the population size of the class. It is easy to see that the calibration equation $\sum_S w_k x_{pk} = N_P$ holds for all p. Moreover, $E_I(N_p / \sum_S a_j x_{pj}) \approx 1$ for a sufficiently large sample because $E_I(\sum_S a_j x_{pj}) = N_p$. Thus, $w_k \approx a_k$. The subscript I denotes expectation treating the I_k as random variables.

Deming and Stephan (1940) extended the notion of poststratification to classes that are not mutually exclusive. Building on the example above, suppose $x_{3k} = 1$ when individual k is of African origin, and $x_{3k} = 0$ otherwise. Their article describes a procedure called *iterative proportional fitting* or *raking* that essentially performs a ratio adjustment for one class at a time, treating the results of the last ratio adjustment as the $\{a_k\}$. The method recycles through the classes as necessary (in practice four or fewer times) until a set of calibration weights is effectively found; that is, the final weights satisfy the calibration equation within roundoff error. It is possible for raking to fail to find a set of final calibration weights, however.

Deming and Stephan called their method "a least squares adjustment," but it is not. Nevertheless, most of the calibration weighting in practice involve a variant of least squares, where the calibration weights have linear the form: $w_k = a_k(1 + c_k \mathbf{x}_k \mathbf{g})$ for some vector, $\mathbf{g}$, and set of constants, $\{c_k \mid k \in S\}$. Deville and Särndal (1992) observed that raking weights have the form:

$$w_k = a_k \exp(\mathbf{x}_k \mathbf{g}).$$

When $\mathbf{x}_k\mathbf{g}$ is small, these weights are asymptotically close to linear calibration weights with all the c_k equal. The authors build on this observation showing how to estimate the model variance and randomization mean squared error simultaneously for an estimator based on calibration weights of the form: $w_k = a_k f(c_k \mathbf{x}_k \mathbf{g})$, where $f(0) = f'(0) = 1$.

Section 2 develops the needed asymptotics for this overview. The general framework follows Isaki and Fuller (1982), but with a stronger focus on the relative mean squared of a calibration estimator. Section 3 discusses the randomization and model-based properties of Särndal, Swensson, and Wretman's (1989) general regression (GREG) estimator, which translates into a calibration estimator with calibration weights in linear form.

Section 4 addresses simultaneous randomization and model-based variance estimation for such estimators. In this, it follows Särndal et al. (1989), Kott (1990), and Valliant (2002). When the first-stage sampling fractions of a multi-stage sample can be ignored, a jackknife procedure is proposed. Its nonstandard replicate weights have convenient generalizations in later sections.

Section 5 proposes a change the definition of calibration weighting. This allows calibration weights to have the form: $w_k = a_k(1 + \mathbf{h}_k \mathbf{g})$, where $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$, as suggested by Estevao and Särndal (2000).

Section 6 reviews nonlinear calibration. With our asymptotic framework, Deville and Särndal's

penetrating insight into variance estimation follows immediately. A generalization of the jackknife procedure from Section 4, although inspired by Deville and Särndal, is new.

Section 7 shows how calibration weighting ideas have been be applied by Fuller, Loughin, and Baker (1994) and Folsom and Singh (2000) for handling unit nonresponse and/or coverage adjustments. In those papers, a quasi-randomization model was assumed, where $1/f(c_k\mathbf{x}_k\gamma)$ was the probability of element k being covered by the frame or responding to the sample. By finding a set of calibration weights such that $w_k = a_k f(c_k\mathbf{x}_k\mathbf{g})$, $\mathbf{g}$ estimates γ implicitly. Here, $c_k\mathbf{x}_k$ will be replaced by the more general $\mathbf{h}_k$. This modest extension allows nonresponse to be a function of some variable(s) of interest (the y_{jk}) while remaining within the calibration framework. Prediction-model unbiasedness is lost. Quasi-randomization consistency is not.

Section 8 offers some concluding remarks ranging from alternative methods for handling unit nonresponse to unresolved issues surrounding sample size.

2. Randomization Consistency and Other Asymptotic Properties

The estimator, t based on a sample of n elements is said to be a consistent estimator for a finite value, T, when $\text{plim}_{n\to\infty}(t) = T$. Fuller (1976, Chapter 5) showed that a sufficient condition for consistency is $\lim_{n\to\infty} E[(t-T)^2] = 0$. This means that both the bias and the mean squared error of t vanishes as the sample size grows arbitrarily large.

For convenience, we focus on a single target variable and assume that all $y_k \geq 0$ and $z_{ak} \geq 0$, where $\mathbf{z}_k = (z_{1k}, ..., z_{Qk})$ is a vector of values associated with element k, and $Q \geq P$. Moreover, we will assume that as the population size, N, and *expected* sample size, n, grow arbitrarily large,

$$0 < L_y \leq \sum_{k\in U} y_k^{\delta}/N \leq B_y < \infty, \quad \delta = 1, ..., 4; \tag{3}$$

$$0 < L_{za} \leq \sum_{k\in U} z_{ak}^{\delta}/N \leq B_{za} < \infty, \quad \delta = 1, ..., 4; \tag{4}$$

for all a, where $(n/N)\pi_k^{-1}$ is one of the components of $\mathbf{z}_k$. Unlike Isaki and Fuller, we are allowing the possibility that N grows at an asymptotically faster rate than n.

Under the regularity conditions, it is not hard to show that when $E(I_jI_k) = \pi_{jk} \leq \pi_j\pi_k$ for $j \neq k$,

$T_y = O(N)$, and

$$\text{Var}_I(t_{y\_E}) = \sum_{j\in U}\sum_{k\in U} [(\pi_{jk} - \pi_j\pi_k)/(\pi_j\pi_k)]y_jy_k$$

$$\leq \sum_{k\in U} [(1/\pi_k) - 1]\, y_k^2 = O(N^2/n),$$

where the last step makes use of Schwartz's inequality (i.e., $\sum y_k^2/\pi_k \leq [\sum y_k^4 \sum 1/\pi_k^2]^{1/2}$). Since the expansion estimator is randomization unbiased, its relative randomization mean squared error is the same as its relative randomization variance, which is O(1/n). Thus, t_{y_E} is randomization consistent with a relative error of $O_P(1/\sqrt{n})$.

The joint selection probabilities in many element sampling plans satisfy $\pi_{jk} \leq \pi_j\pi_k$ whenever $j \neq k$. Simple random sampling, stratified simple random sampling, and Poisson sampling are among them. Asok and Sukhatme (1976) showed that $\pi_{jk} = [(n-1)/n]\pi_j\pi_k[1 + O(n/N)]$ under Sampford sampling and Goodman-Kish sampling. Consequently, both sampling plans are in this class as well for sufficiently large N when $O(N) \geq O(n^{3/2})$,

In many multi-stage sampling plans, when j and k are in the same primary sampling unit (PSU), π_{jk} will usually exceed $\pi_j\pi_k$. To extend asymptotic properties to multi-stage samples where $\pi_{jk} \leq \pi_j\pi_k$ need hold only when j and k are in *different* PSUs, we first divide the population into PSUs, and assume that the number of these PSUs, N_1, grows proportionally with N. We similarly assume that the expected number of PSUs in the first-stage sample, n_1, grows proportionally with n. We add the assumption that the individual population size for each PSU i is bounded. Finally, we replace equations (3) and (4) with PSU-level analogues, letting, for example, $t_{y(i)}$ be the sum of the y-values across all then elements in i. Equation (3) can be replaced by $0 < L_{y'} \leq \sum t_{y(i)}^{\delta}/N_1 \leq B_{y'} < \infty$, where the summation is over the N_1 PSUs. The proof is left to the reader who should note that $\pi_{jk} \leq \max\{\pi_j, \pi_k\}$, which implies $(\pi_{jk} - \pi_j\pi_k)/(\pi_j\pi_k) \leq \max\{1/\pi_j, 1/\pi_k\} - 1$.

One common sampling plan that does *not* lead to randomization consistent estimation is systematic sampling from an ordered list. The problem is that given any element k, the number of other elements j such that $\pi_{jk} > \pi_j\pi_k$ grows at the same rate as the (expected) sample size.

3. The General Regression Estimator

Due to the popularity of the book, *Model-Assisted Survey Sampling* (Särndal, Swensson, and Wretman, 1992), it is common to call the randomization-consistent regression estimator the "general regression" or "GREG estimator." For our purposes, it has the form:

$$t_{y\_GREG} =$$

$$t_{y\_E} + (T_{\mathbf{x}} - \sum a_k\mathbf{x}_k)(\sum c_ka_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum c_ka_k\mathbf{x}_k'y_k, \tag{5}$$

$$k\in S \qquad k\in S \qquad\qquad k\in S$$

where $\mathbf{x}_k$ is a row vector composed of components of $\mathbf{z}_k$ in equation (4), $a_k = 1/\pi_k$ for $k \in S$ (as before), c_k is also a component of $\mathbf{z}_k$, which may or may not be a function of $\mathbf{x}_k$, and $\lim_{N\to\infty} \sum_U c_k\mathbf{x}_k'\mathbf{x}_k/N = \mathbf{\Phi}$ is positive definite matrix. This last condition means that $\sum_S c_k a_k \mathbf{x}_k'\mathbf{x}_k$ will usually be invertible in practice. We will assume that it is always invertible for convenience.

Sometimes the c_k within equation (5) are assumed to be proportional to the inverses of $E(\epsilon_k^2)$. We do not make that assumption here.

Let $\mathbf{b} = (\sum_S c_k a_k \mathbf{x}_k'\mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1}\mathbf{x}_k'y_k$, and $\mathbf{B} = (\sum_U c_k\mathbf{x}_k'\mathbf{x}_k)^{-1} \sum_U c_k^{-1}\mathbf{x}_k'y_k$. The GREG estimator can be written as $t_{y\_GREG} = t_{y\_E} + (T_\mathbf{x} - \sum_{k\in S} a_k\mathbf{x}_k)\mathbf{b}$, which is close to the pseudo-difference estimator:

$$t_{y\_PDIF} = t_{y\_E} + (\sum_{k\in U} \mathbf{x}_k\mathbf{B} - \sum_{k\in S} a_k\mathbf{x}_k\mathbf{B}), \qquad (6)$$

where $\mathbf{x}_k\mathbf{B}$ plays the role of x_k in the standard difference estimator. The pseudo-difference estimator is randomization unbiased.

The GREG estimator in equation (5) can be rewritten in calibration form as $t_{y\_GREG} = \sum_S w_k y_k$, where

$$w_k = a_k + (T_\mathbf{x} - \sum_{j\in S} a_j\mathbf{x}_j)(\sum_{j\in S} c_j a_j \mathbf{x}_j'\mathbf{x}_j)^{-1} c_k a_k \mathbf{x}_k'. \qquad (7)$$

Strictly speaking, the w_k are functions of the realized sample, S, and the $c_k a_k$, but we suppress that in the notation for convenience.

3.1 *The Randomization-Based Properties of the GREG Estimator*

Let us assume that the regularity conditions and sample plan are such that $t_{y\_E} - T_y = O_P(N/\sqrt{n})$, $\sum_S a_k\mathbf{x}_k - T_\mathbf{x} = \mathbf{O}_P(N/\sqrt{n})$, and $\sum_k c_k a_k \mathbf{x}_k'\mathbf{f}_k - \sum_U c_k\mathbf{x}_k'\mathbf{f}_k = \mathbf{O}_P(N/\sqrt{n})$, where $\mathbf{f}_k$ can be $\mathbf{x}_k$ or y_k. Letting $e_k = y_k - \mathbf{x}_k\mathbf{B}$, so that $\sum_U c_i\mathbf{x}_i'e_i = 0$, and $\sum_S c_k a_k \mathbf{x}_k'e_k = \mathbf{O}_P(N/\sqrt{n})$. We can express the error of t_{y_GREG} as

$$t_{y\_GREG} - T_y = \sum_{k\in S} w_k y_k - \sum_{k\in U} y_k = \sum_{k\in S} w_k e_k - \sum_{k\in U} e_k$$

$$= \sum_{k\in S} a_k e_k + (T_\mathbf{x} - \sum_{k\in S} a_k\mathbf{x}_k)(\sum_{k\in S} c_k a_k \mathbf{x}_k'\mathbf{x}_k)^{-1} \sum_{k\in S} c_k a_k \mathbf{x}_k'e_k - \sum_{k\in U} e_k$$

$$= \sum_{k\in S} a_k e_k - \sum_{k\in U} e_k + O_p(N/n). \qquad (8)$$

Since $e_k \le y_k + |\mathbf{x}_k\mathbf{B}|$, it is not hard to see the GREG estimator is randomization consistent with a relative randomization bias and mean squared error of asymptotic order 1/n. The randomization bias is an asymptotically insignificant contributor to the mean squared error, mse, when $\text{plim}_{n\to\infty}(n\,\text{mse}/N^2) > 0$, a mild condition violated when nearly all the e_k in the population are zero, which we assume not to be the case.

3.2 *Model-based Properties of the GREG Estimator*

Suppose the y_k are random variables that satisfy the linear model in equation (2). In addition, assume $E(\epsilon_k | \{\mathbf{x}_g | g\in S\}, \{I_g | g\in U\}) = E(\epsilon_k\epsilon_j | \{\mathbf{x}_g | g\in S\}, \{I_g | g\in U\}) = 0$ for $k \ne j$, and $E(\epsilon_k^2 | \mathbf{x}_k, I_k) = \sigma_k^2$. The σ_k^2 *need not be known*. Moreover, there is no reason that I_k cannot be a function of the components of $\mathbf{z}_k$.

It is easy to see that as long as the regression weights satisfy the calibration equation, $\sum_S w_k\mathbf{x}_k = T_\mathbf{x}$, t_{y_GREG} will be model unbiased. Its model variance, as well as the model variance of any calibration estimator, is

$$E_\epsilon[(t_{y\_GREG} - T_y)^2]$$

$$= E_\epsilon[(\sum_{k\in S} w_k y_k - \sum_{k\in U} y_k)^2]$$

$$= E_\epsilon[(\sum_{k\in S} w_k\epsilon_k - \sum_{k\in U} \epsilon_k)^2]$$

$$= \sum_{k\in S} w_k^2\sigma_k^2 - 2\sum_{k\in S} w_k\sigma_k^2 + \sum_{k\in U} \sigma_k^2$$

$$= \sum_{k\in S} w_k^2\sigma_k^2 - \sum_{k\in S} w_k\sigma_k^2 - (\sum_{k\in S} w_k\sigma_k^2 - \sum_{k\in U} \sigma_k^2)$$

$$= \sum_{k\in S} w_k^2\sigma_k^2 - \sum_{k\in S} w_k\sigma_k^2 + O_P(N/\sqrt{n}), \qquad (9)$$

under mild condition, in particular, those where $w_k = a_k[1 + O_p(1/\sqrt{n})]$, and $\sum_S a_k\sigma_k^2 - \sum \sigma_k^2 = O_P(N/\sqrt{n})$. Notice that we are using randomization-based asymptotic results in a purely model-based context. We are not, however, averaging over all possible samples, which is what randomization-based theory routinely does.

When σ_k^2 has the form $\mathbf{x}_k\zeta$, for some not-necessarily-specified vector ζ, then $\sum_S w_k\sigma_k^2 = \sum_U \sigma_k^2$, and the model varianceof t_{y_GREG} collapses to $\text{Var}_\epsilon(t_{y\_GREG}) = \sum_S (w_k^2 - w_k)\sigma_k^2$ exactly. Alternatively, when $N \ge O(n^{3/2})$, the model variance is dominated by $\sum_S w_k^2\sigma_k^2$ if $\pi_k = O(n/N)$.

For a multi-stage sample it makes sense to allow the possibility that ϵ_k and ϵ_j are correlated when

k and j are in the same PSU, but not otherwise. Under the regularity conditions discussed previously for a multi-stage sample, if $\pi_{jk} \le \pi_j\pi_k$ for j and k from *different* PSUs and $N \ge O(n^2)$, it is not hard to show that the model variance of the GREG estimator is dominating by $\sum_{i\in S'} E_\epsilon[(\sum_{k\in S(i)} w_k\epsilon_k)^2]$, where S(i) is the set of sampled elements in PSU i and S' is the set of PSUs selected for the sample.

Let us return to the model with no correlation among the elements. The model variance of t_{y_GREG} is $O_P(N^2/n)$ under mild conditions we assume to hold. If we are willing to drop $O_P(N^2/n_E^{3/2})$ terms (so that $w_k \approx 1/\pi_k$, and $\sum_{k\in S} a_k\sigma_k^2 - \sum_{k\in U}\sigma_k^2 \approx 0$), the model variance of t_{y_GREG} can be approximated by $E_\epsilon[(t_{y\_GREG} - T)^2] \approx \sum_{k\in S}(\sigma_k^2/\pi_k^2)(1-\pi_k)$.

The randomization expectation of the model variance of t_R is then

$$E_p\{E_\epsilon[(t_{y\_GREG} - T)^2]\} \approx \sum_{k\in U} (\sigma_k^2/\pi_k)(1-\pi_k). \quad (10)$$

The right hand side of equation (10) was called *anticipated variance* of the GREG by Isaki and Fuller (1982), although the equation goes back considerably further in the literature and "anticipated mean squared error" would have been better. They used it to mean $E_\epsilon\{E_p[(t_{y\_GREG} - T_y)^2]\}$, what that model anticipates the randomization mean squared error to be. The expectation operators can be switched when ϵ_k and ϵ_k^2 are uncorrelated with I_k given $\mathbf{z}_k$.

Notice that the joint selection probabilities have no effect on the asymptotic anticipated variance expressed by the right hand side of equation (10). Similarly, the choice for c_k does not matter in this context.

3.3 *Some Choices for the c_k*

If the choice for c_k has no effect on the anticipated variance of a GREG estimator, then how should one choose the c-values in practice? One popular choice is to set all the $c_k = 1$ when some component or linear combination of components of $\mathbf{x}_k$ is 1. Alternatively, when $\mathbf{x}_k$ is a scalar or only one component of $\mathbf{x}_k$ is positive while the rest are zero, it is popular to set $c_k = 1 / \max\{x_{1k}, ..., x_{Pk}\}$. Under either of these two strategies, there is a vector λ such that $c_k\mathbf{x}_k\lambda = 1$ for all $k \in S$. As a consequence, the GREG can be put into projection form: $t_{y\_GREG} = T_x\mathbf{b}$, where $\mathbf{b} = (\sum_S c_ka_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_ka_k\mathbf{x}_k'y_k$. This is because

$$\begin{aligned}\sum_S a_k\mathbf{x}_k\mathbf{b} &= \sum_S c_ka_k\lambda'\mathbf{x}_k'\mathbf{x}_k\mathbf{b}\\ &= \lambda'(\sum_S c_ka_k\mathbf{x}_k'\mathbf{x}_k)\mathbf{b}\\ &= \lambda'(\sum_S c_ka_k\mathbf{x}_k'\mathbf{x}_k)(\sum_S c_ka_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_ka_k\mathbf{x}_k'y_k\\ &= \lambda'\sum_S c_ka_k\mathbf{x}_k'y_k = \sum_S a_ky_k,\end{aligned}$$

so $t_{y\_GREG} = \sum_S a_ky_k + (T_x - \sum_S a_k\mathbf{x}_k)\mathbf{b} = T_x\mathbf{b}$.

Using similar reasoning, Brewer (1994) showed that when one sets $c_k\mathbf{x}_k\lambda = (1 - \pi_k)$ for all $k \in S$ instead of $c_k\mathbf{x}_k\lambda = 1$, t_{y_GREG} can be put in prediction form: $t_{y\_GREG} = \sum_S y_k + (T_x - \sum_S a_k\mathbf{x}_k)\mathbf{b}$. More importantly, perhaps, this alternative choice for the c_k results in a set of *cosmetically calibrated* weights that are empirically less likely to have a member smaller than 1. Brewer advised bounding calibration weights from below by 1, so that each element at least represents itself in the estimate $t_{y\_CAL} = \sum_S w_ky_k$.

The key to cosmetic calibration's empirical success in this regard – which is not absolute – appears to be that when the sampling weight, a_k, is close to 1, cosmetic calibration will not allow it to change by very much. Indeed, when $a_k = 1$, then $c_k = 0$, and $w_k = 1$.

4. Variance Estimation

If model in equation (2) holds, and the element errors are uncorrelated with $E(\epsilon_k^2) = \sigma_k^2$, then equation (9) tells us that under certain conditions, the model variance of an estimator in calibration form is (approximately) $V_m = \sum_S (w_k^2 - w_k)\sigma_k^2$. This suggests the following estimator for model variance:

$$v_m = \sum_{k\in S} (w_k^2 - w_k)r_k^2, \quad (11)$$

where $r_k = y_k - x_k\mathbf{b}$ is a sample residual, and $\mathbf{b}$ is *any* model-unbiased estimator for the model parameter, β. Under mild assumptions similar to the regularity conditions in equation (4), $E_\epsilon(r_k^2) = \sigma_k^2 + O_P(1/n)$, and $E_\epsilon(v_m) = V_m[1 + O_p(1/n)]$.

From equation (8), we can conclude that randomization mean squared error of the estimator *under Poisson sampling* is approximately $V = \sum_{k\in U}[a_k - 1]e_k^2$. If $w_k = a_k[1 + O_P(1/\sqrt{n_E})]$, then v_m is a reasonable mean-squared-error estimator when $r_k^2 \approx e_k^2$. Let $r_k = y_k - \mathbf{x}_k\mathbf{b}$ and $e_k = y_k - \mathbf{x}_k\mathbf{B}$, where

$\mathbf{b} = (\sum_{k\in S} c_ka_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_{k\in S} c_ka_k\mathbf{x}_k'y_k$, and
$\mathbf{B} = (\sum_{k\in U} c_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_{k\in U} c_k^{-1}\mathbf{x}_k'y_k$.

Since $\mathbf{b} = \mathbf{B}[1 + O_P(1/\sqrt{n})]$ under the regularity conditions in equation (4), $r_k^2 = e_k^2 + O_P(1/\sqrt{n})$.

Särndal et al. (1989) proposed this variance/mean-squared-error estimator for the GREG under an arbitrary sampling plan

$$v_{SSW} = \sum_{k\in S}\sum_{j\in S} [(\pi_{kj} - \pi_k\pi_j)/\pi_{kj}](w_kr_k)(w_jr_j). \quad (12)$$

Developing asymptotic properties for v_{SSW} can be elusive when it contains $n(n-1)/2$ distinct terms. That is

not a problem under stratified simple random sampling, where

$$v_{ST1} = \sum_{\alpha=1}^{A} (n_\alpha/[n_\alpha - 1]) \sum_{k\in S_\alpha} (1 - n_\alpha/N_\alpha) (w_k r_k - \sum_{j\in S_\alpha} w_j r_j /n_\alpha)^2, \quad (13)$$

S_α denotes the sample of n_α units in stratum α (α = 1, ..., A), and U_α the stratum population containing N_α elements.

Let us assume the same model assumptions and regularity conditions for a multi-stage sample as before and that $N \geq O(n^2)$. The model variance of a calibration estimator is then approximately
$V_m = \sum_{i\in S'} E_\epsilon [(\sum_{k\in S(i)} w_k \epsilon_k)^2]$, where S(i) is the set of sampled elements in PSU i, and S' is the set of PSUs selected in the first stage of sampling.

Consider a GREG estimator under stratified multi-stage sampling, where $\pi_{kj} \leq \pi_k \pi_j$ for k and j from different PSUs, and the first-stage selection probabilities are ignorably small. The following variance estimator has good randomization and model-based properties under mild conditions:

$$v_{ST2} = \sum_{\alpha=1}^{A} (n_{1\alpha}/[n_{1\alpha} - 1]) \{ \sum_{j\in S_{1\alpha}} (\sum_{k\in S_{\alpha j}} w_k r_k)^2 - (\sum_{j\in S_{1\alpha}} \sum_{k\in S_{\alpha j}} w_k r_k)^2 /n_\alpha \}, \quad (14)$$

where α denotes a first-stage stratum of PSU's, $n_{1\alpha}$ the number of sampled PSU's in stratum α, $S_{1\alpha}$ the set of sampled PSU's in α, and S_{α_j} the set of subsampled elements from PSU j of stratum α. There can be many stages of sampling involved.

It is not hard to show that v_{ST2} is asymptotically indistinguishable from the jackknife variance estimator:

$$v_J = \sum_{\alpha=1}^{A} ([n_\alpha - 1]/n_\alpha) \{ \sum_{j\in S_{1\alpha}} (t_{y\_CAL(\alpha j)} - t_{y\_CAL})^2 \}, \quad (15)$$

where $t_{y\_CAL(\alpha j)} = \sum_{k\in S} w_{k(\alpha j)} y_k$, and the *jackknife replicate calibration weights* are

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)}/a_k + (\sum_{m\in U} \mathbf{x}_m - \sum_{m\in S} w_m [a_{m(\alpha j)}/a_m] \mathbf{x}_m) (\sum_{m\in S} a_{m(\alpha j)} c_m \mathbf{x}_m' \mathbf{x}_m)^{-1} a_{k(\alpha j)} c_k \mathbf{x}_k',$$

where
$a_{k(\alpha j)}$ = 0 when k is in PSU j of stratum h,
$a_{k(\alpha j)}$ = a_k when k is not in stratum α at all, and
$a_{k(\alpha j)} = (n_\alpha/[n_\alpha - 1])a_k$ otherwise.
The $w_{k(\alpha j)}$ are constrained so that $\sum_{k\in S} w_{k(\alpha j)} \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k$ for all αj. Now, under our assumptions,

$$\sum_{m\in U} \mathbf{x}_m - \sum_{m\in S} w_m [a_{m(\alpha j)}/a_m] \mathbf{x}_m =$$

$$(n_\alpha/[n_\alpha - 1])(\sum_{k\in S(\alpha j)} w_k \mathbf{x}_k - \sum_{k\in S(\alpha)} w_k \mathbf{x}_k /n_\alpha) = \mathbf{O}_P(N/n),$$

$$\sum_{m\in S} c_m a_{m(\alpha j)} \mathbf{x}_m' \mathbf{x}_m) = \mathbf{O}_P(N),$$

and

$$\sum_{m\in S} c_m a_{m(\alpha j)} \mathbf{x}_m' e_m = \mathbf{O}_P(N/\sqrt{n}),$$

where S(α) is the set of elements in stratum α, and S(αj) is the set of elements in PSU j of stratum α. As a result,

$$t_{y\_CAL(\alpha j)} - t_{y\_CAL} = \sum_{k\in S} w_{k(\alpha j)} e_k - \sum_{k\in S} w_k e_k = (n_\alpha/[n_\alpha - 1])(\sum_{k\in S(\alpha)} w_k e_k /n_\alpha - \sum_{k\in S(\alpha j)} w_k e_k) + O_P(N/n^{3/2}),$$

and

$$v_J = v_{ST2} [1 + O_P(1/\sqrt{n})] \text{ when } \text{plim}_{n\to\infty}(n v_{ST2}/N^2) > 0.$$

The replicate weights described above are nonstandard. More common is

$$w_{k(\alpha j)} = a_{k(\alpha j)} + (\sum_{m\in U} \mathbf{x}_m - \sum_{m\in S} a_{m(\alpha j)} \mathbf{x}_m)(\sum_{m\in S} c_m a_{m(\alpha j)} \mathbf{x}_m' \mathbf{x}_m)^{-1} c_k a_{k(\alpha j)} \mathbf{x}_k',$$

which "look like" the original calibration weights. Our version generates a v_J with a model expectation closer to $\sum_{i\in S'} E_\epsilon [(\sum_{k\in S(i)} w_k \epsilon_k)^2]$. Replacing e_k in the arguments above by ϵ_k, it is not hard to show that $E_\epsilon(v_J) = V_m [1 + O_P(1/n)]$ under mild conditions.

5. Redefining Calibration Weights

In their original definition of calibration weights, Deville and Särndal (1992) required that the set of calibration weights, $\{w_k \mid k\in S\}$ minimize some distance function between the members of the set and the original sampling weights, the a_k, subject to satisfying the calibration equation. As a result, the calibration estimator, $t_{y\_CAL} = \sum_S w_k y_k$, was both unbiased under the model in equation (2) and usually randomization consistent.

Estevao and Särndal (2002) suggested removing the requirement that the calibration weights minimize a distance function. Instead, they essentially proposed that the w_k need only satisfy the calibration equation and be of the "functional form:"

$$w_k = a_k(1 + \mathbf{h}_k \mathbf{g}), \quad (16)$$

where $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$ such that $\sum_S a_k \mathbf{h}_k' \mathbf{x}_k$ is invertible, and $\mathbf{g}$ is a column vector of that same dimension. It is a generalization of the GREG where $\mathbf{h}_k$ effectively replaces $c_k \mathbf{x}_k$

It is not hard to see that

$\mathbf{g} = (\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1} (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)'$.

Moreover, if the components of $\mathbf{h}_k$ are components of $\mathbf{z}_k$ in equation (4), the regularity conditions hold, and $\sum_S a_j \mathbf{h}_j' \mathbf{x}_j / N$ is invertible both for the realized N and in the probability limit, then

$$t_{y\_CAL} = \sum_S w_k y_k = \sum_S a_k y_k + (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1} \sum_S a_k \mathbf{h}_k' y_k$$

is randomization consistent whenever t_{y_E} is. It is unbiased under the linear prediction model in equation (2) when $E(\epsilon_k \mid \{\mathbf{x}_g, \mathbf{h}_g \mid g \in S\}, \{I_g \mid g \in U\}) = 0$ for all $k \in U$.

This suggest another alternative definition of calibration weights: a set of weights, $\{w_k \mid k \in S\}$, such that,

1) the w_k satisfy the calibration equation for $\{\mathbf{x}_k \mid k \in U\}$, and

2), $t_{y\_CALC} = \sum_S w_k y_k$ is randomization consistent whenever t_{y_E} is under mild conditions.

That is the definition we will use.

It follows that Estevao and Särdnal's functional-form calibration is indeed a form a calibration weighting. Borrowing from econometric theory, Kott (2003) called the components of $\mathbf{h}_k$ that were not linear combinations of components of $\mathbf{x}_k$ "instrumental variables." Both Kott and Estevao and Särdnal discussed choices for the $\mathbf{h}_k$ that may decrease the likelihood of some calibration weights being less than unity.

Space limitations prevent us from seeing how Rao's randomization-optimal estimator (1994), a regression estimator that asymptotically minimizes randomization mean squared error, can be put into calibration form. Breidt and Opsomer (2000) show how a randomization-consistent estimator incor-porating local polynomial regression can also be put into calibration form.

6. Nonlinear Calibration

Building on ideas in Deville and Särndal (1992), we can generalize the linear form for the calibration weights in equation (15) to

$$w_{k\_GEN} = a_k[1 + f(\mathbf{h}_k \mathbf{g}^*)], \qquad (19)$$

where f is a monotonic, twice-differentiable function with $f(0) = 0$, $f'(0) = 1$ ($f'(0)$ is the first derivative of f evaluated at 0), and $\mathbf{g}^*$ is chosen so that the calibration equation holds. This formulation of f(.) is different than that in the literature and the introduction for convenience. An extension of equation (19) with potentially different f(.) across the sampled elements is straightforward and left to the reader.

A solution, $\mathbf{g}^*$, to equation (19) can be approached iteratively. One can start with $\mathbf{g}^{(0)} = \mathbf{0}$; that is, $\sum_S w_k^{(0)} y_k$, where $w_k^{(0)} = a_k[1 + f(0)] = a_k$. For r = 1, 2, ..., one then sets

$$\mathbf{g}^{(r)} = \mathbf{g}^{(r-1)} + [\sum_S f'(\mathbf{h}_k \mathbf{g}^{(r-1)}) a_k \mathbf{h}_k' \mathbf{x}_k]^{-1} (T_\mathbf{x} - \sum_S w_k^{(r-1)} \mathbf{x}_k)' .$$

Note that $\mathbf{g}^{(1)}$ equals the $\mathbf{g}$ in $w_{k\_LIN} = a_k(1 + \mathbf{h}_k \mathbf{g})$. A Taylor expansion around zero reveals $f(\mathbf{h}_k \mathbf{g}^{(1)}) = \mathbf{h}_k \mathbf{g}^{(1)} + O_p(1/n_E)$ under our usual regularity conditions, so $\sum_S w_k^{(1)} y_k = \sum_S w_{k\_LIN}\, y_k + O_P(N/n) = T_y[1 + O_P(1/n)]$. Furthermore, it is not difficult to see that $w_{k\_GEN} = w_{k\_LIN}[1 + O_P(1/n)]$, an equality that proves helpful in variance estimation. One should be aware, however, that there may not be a set of weights that both can be expressed in the form of equation (19) and satisfy calibration equation.

The most common example in practice of a nonlinear f is $f(\mathbf{h}_k \mathbf{g}) = \exp(\mathbf{x}_k \mathbf{g}) - 1$, where the values of each of the components of $\mathbf{x}_k$, denoted $x_{1k}, ..., x_{Pk}$, are either 0 or 1. That is effectively the form of Deming and Stephan's calibration weights computed via iterative proportional fitting. Many have observed that the iterative routine described above can be used even when the components of $\mathbf{x}_k$ are not binary. Moreover, the calibration weights produced (when a set is produced) are always nonnegative.

Returning to the general case, since $w_{k\_GEN} = w_{k\_LIN}[1 + O_P(1/n)]$ under conditions we assume to hold, it is not hard to show that the variance estimators in Section 4 apply equally well to the calibration estimator based on the w_{k_GEN} when $r_k = y_k - \mathbf{x}_k \mathbf{b}_{INST}$, and $\mathbf{b}_{INST} = (\sum a_k \mathbf{h}_k' \mathbf{x}_k)^{-1} \sum_S a_k \mathbf{h}_k' y_k$. This is a mild generalization of Deville and Särndal's insight replacing their $c_k \mathbf{x}_k$ by $\mathbf{h}_k$. Following the logic of their article, one would also replace the a_k in our $\mathbf{b}_{INST}$ by w_{k_LIN}. That isn't *wrong* – the two versions of r_k are within $O_p(1/\sqrt{n})$ of each other, but there is little reason for doing what the authors suggest.

Deville and Särndal's insight extends further. For the jackknife variance estimator in equation (15), the jackknife replicate calibration weights, the $w_{k(\alpha j)}$, can be computed like they were in Section 4 with $\mathbf{h}_k'$ replacing $c_k \mathbf{x}_k'$; that is,

$$w_{k(\alpha j)} = w_k\, a_{k(\alpha j)}/a_k + (\sum_{m \in U} \mathbf{x}_m - \sum_{m \in S} w_m [a_{m(\alpha j)}/a_m] \mathbf{x}_m) (\sum_{m \in S} a_{m(\alpha j)} \mathbf{h}_m' \mathbf{x}_m)^{-1} a_{k(\alpha j)} \mathbf{h}_k'.$$

This simplifies their computation in practice since iteration is not required.

7. Using Calibration to Adjust for Nonresponse (or Undercoverage)

One popular way of handling unit (whole-element) nonresponse is to treat response as an additional phase of Poisson sampling. This is the essence of a *quasi-*

randomization model. Each element k in the original sample, now denoted S', is assumed to have a probability of response p_k, and the probability of jointly "choosing" elements k and j is $p_k p_j$. Furthermore, the value of p_k is independent of whether k is chosen for the original sample. It is often possible to construct a set of weights such that are randomization consistent under the quasi-randomization model.

We are interested here in a particular way of constructing those weights. To this end, we assume that the quasi-randomization model is correct. Each element has attached to it a row vector of auxiliary variables, $\mathbf{x}_k$, for which $T_{\mathbf{x}} = \sum_U \mathbf{x}_j$ is known. Finally, each p_k is assumed to have the form:

$$p_k = 1/[1 + f(\mathbf{h}_k \gamma)], \tag{20}$$

where γ is unknown, $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$, and $\sum_S a_k \mathbf{h}_k' \mathbf{x}_k /N$, where S now denotes the "subsample" of respondents, is invertible both for the realized N and in the probability limit. The function f is assumed to be monotonic and twice differentiable. Its functional form is known, but the value of the governing parameter, γ, is not.

The most obvious choice for $\mathbf{h}_k$ when postulating the response model in equation (20) is $\mathbf{x}_k$ itself. In some applications, however, some component(s) of $\mathbf{x}_k$ may have been chosen because it was the best measures we had for a variable *before* sampling. An example of such a variable in a survey of farms is the total land area of an operation. After collecting survey values, it may be possible to replace a component of $\mathbf{x}_k$ (in $\mathbf{h}_k$) with a better measure of the variable in question. In our example, response is more likely a function of the actual land area of a farm than a predetermined proxy for that value. As a result, replacing the corresponding proxy value with the survey value is tempting. A theoretical problem with this procedure is discussed below.

Using the iterative method described in the last section to find $\mathbf{g}^*$, we will often be able to uncover a row vector, $\mathbf{g}$, such that $T_{\mathbf{x}} = \sum_S a_k[1 + f(\mathbf{h}_k \mathbf{g})]\mathbf{x}_i$. As a result, estimating T_y with $t_{y\_CAL} = \sum_S w_k y_k$, where the adjusted calibration weights have the form,

$$w_k = a_k[1 + f(\mathbf{h}_k \mathbf{g})],$$

may have good properties under the linear prediction model:

$$y_k = \mathbf{x}_k \beta + \epsilon_k,$$

where $E(\epsilon_k \mid \{\mathbf{x}_g, \mathbf{h}_g \mid g \in S\}, \{ I_g \mid g \in U\}) = 0$ for all $k \in U$, $I_k = 1$ if element k is both in the original sample and responds, 0 otherwise.

Prediction-model unbiasedness is simply a result of the weights satisfying the calibration equation (the prefix "prediction" to needed to distinguish this model from the quasi-random one). Note, however, that if some components of $\mathbf{h}_k$ come from the survey rather than $\mathbf{x}_k$, the prediction-model assumption that $E(\epsilon_k \mid \mathbf{h}_k) = 0$ can be problematic. At the extreme, consider the case where one such component is y_k itself. Obviously, $E(\epsilon_k \mid y_k)$ is not usually 0. In the example described above, y_k may be the total land area on farm operation k. Putting total land area in $\mathbf{h}_k$ makes the associated calibration estimator prediction-model biased.

Whether or not t_{y_CAL} can reasonably be called prediction-model unbiased has no effect on its quasi-randomization-based properties. Since $T_{\mathbf{x}} = \sum_S a_k[1 + f(\mathbf{h}_k \mathbf{g})]\mathbf{x}_i$, our assumptions and the mean value theorem reveal

$$T_{\mathbf{x}} - \sum_S a_k[1 + f(\mathbf{h}_k \gamma)]\mathbf{x}_k = - \sum_S a_k [f'(\mathbf{h}_k \mathbf{g}^o)\mathbf{h}_k (\mathbf{g} - \gamma)]\mathbf{x}_k$$
$$= \mathbf{O}_p(N/\sqrt{n})$$

for some $\mathbf{h}_k \mathbf{g}^o$ between $\mathbf{h}_k \mathbf{g}$ and $\mathbf{h}_k \gamma$. From this we see that if $\sum_S a_j f'(\mathbf{h}_j \gamma)_j \mathbf{h}_j' \mathbf{x}_j /N$ is invertible both for the realized N and at the probability limit (recall that f is monotonic so f' is never zero), then

$$g - \gamma = - \{ \sum_S a_j f'(\mathbf{h}_j \mathbf{g}^0)_j \mathbf{h}_j' \mathbf{x}_j \}^{-1} \{T_{\mathbf{x}} - \sum_S a_i[1 + f(\mathbf{h}_i \gamma)]\mathbf{x}_i \}$$
$$= - \{ \sum_S a_j f'(\mathbf{h}_j \gamma)_j \mathbf{h}_j' \mathbf{x}_j \}^{-1} \{T_{\mathbf{x}} - \sum_S a_i[1 + f(\mathbf{h}_i \gamma)]\mathbf{x}_i \} + \mathbf{O}_p(1/n)$$
.

The estimator t_{y_CAL} has an error of

$$t_{y\_CAL} - T_y = \sum_S a_k[1 + f(\mathbf{h}_k \mathbf{g})]y_k - \sum_U y_k$$
$$= \sum a_k[1 + f(\mathbf{h}_k \mathbf{g})]e_k - \sum e_k ,$$

where $e_k = y_k - \mathbf{x}_k (\sum_U f'(\mathbf{h}_j \gamma) p_j \mathbf{h}_j' \mathbf{x}_j)^{-1} \sum_U f'(\mathbf{h}_j \gamma) p_j \mathbf{h}_j' y_j$, and $p_j = 1/[1 + f(\mathbf{h}_j \gamma)]$ so $\sum_S a_k f'(\mathbf{h}_k \gamma) \mathbf{h}_k' e_k = \mathbf{O}_p(N/\sqrt{n})$. Continuing:

$$t_{y\_CAL} - T_y = \sum_{k \in S} a_k[1 + f(\mathbf{h}_k \gamma)]e_k - \sum_{k \in U} e_k + \sum_{k \in S} a_k \{ f(\mathbf{h}_k \mathbf{g}) - f(\mathbf{h}_k \gamma) \} e_k$$
$$= \sum a_k[1 + f(\mathbf{h}_k \gamma)]e_k - \sum e_k + \sum a_k f'(\mathbf{h}_k \gamma) \mathbf{h}_k (\mathbf{g} - \gamma) e_k + O_p(N/n)$$
$$= \sum a_k[1 + f(\mathbf{h}_k \gamma)]e_k - \sum e_k + (\mathbf{g} - \gamma)' \sum a_k f'(\mathbf{h}_k \gamma) \mathbf{h}_k' e_k + O_p(N/n)$$
$$= \sum a_k[1 + f(\mathbf{h}_k \gamma)]e_k - \sum e_k + O_p(N/n)$$

$$k\in S \qquad k\in U \qquad (21)$$

Thus, t_{y_CAL} is quasi-randomization consistent under mild conditions whenever $t = \sum_S a_k[1 + f(\mathbf{h}_k \gamma)]y_k$ is.

To estimate the quasi-randomization mean squared error of t_{y_CAL} (i.e., the estimator's randomization mean squared error under the quasi-randomization model), we first note that the probability that elements k and j, $k \neq j$, are both in the respondent subsample is $\pi_{kj}^* = \pi_{kj}p_kp_j$. Let $\pi_k^* = \pi_k p_k$, and recall that $a_k = 1/\pi_k$ and $1/p_k = a_k[1 + f(\mathbf{h}_k \gamma)]$. From equation (21), we see that the randomization mean squared error of t_{y_CAL} is approximately

$$E_I[(t_{y\_CAL} - T_y)^2] \approx \sum_{k\in U}\sum_{j\in U}(\pi_{kj}^* - \pi_k^*\pi_j^*)(e_k/\pi_k^*)(e_j/\pi_j^*)$$

$$= \sum_{k\in U}(1 - \pi_k^*)e_k^2/\pi_k^* + \sum_{k\in U}\sum_{j\in U, k\neq j}(\pi_{kj} - \pi_k\pi_j)(e_k/\pi_k)(e_j/\pi_j) \qquad (22)$$

If the original sample is Poisson, then v_m in equation (11) with

$$r_k = y_k - \mathbf{x}_k\Big[\sum_{j\in S} a_j f'(\mathbf{h}_j \mathbf{g})\,\mathbf{h}_j'\mathbf{x}_j\Big]^{-1}\sum_{j\in S} a_j f'(\mathbf{h}_j \mathbf{g})\mathbf{h}_j'y_j, \qquad (23)$$

serves as both a reasonable estimator for prediction-model variance and quasi-randomization mean squared error under mild conditions, since $w_k \approx 1/\pi_k^*$ and $r_k \approx e_k$. A close relative of the non-intuitive sample residual in equation (23) can be found in Folsom and Singh (2000).

For a general design, we can get close to the a good variance/mean-squared-error estimator by starting with v_{SSW} in equation (12), where r_k is again defined by equation (23). We need to add a term like

$$v_{add} = \sum_{k\in S}(w_k^2\pi_k - w_k)r_k^2,$$

so that $\sum_U(1 - \pi_k^*)e_k^2/\pi_k^*$ in equation (22) is estimated by $\sum_S(w_k^2 - w_k)r_k^2$ rather than $\sum_S w_k^2(1 - \pi_k)r_k^2$. This correction to v_{SSW} in equation (12) has good prediction-model-based properties when the ϵ_k are uncorrelated, and $\sigma_k^2 = \mathbf{x}_k\zeta$, for some ζ. It can be made even the in the absence of nonresponse.

When the actual sample is multistage, and the first stage selection probabilities are ignorably small, v_{ST2} in equation (14) can be used as the variance/mean-squared-error estimator with r_k defined once more by equation (23).

Observe that when there is no nonresponse, $\gamma = 0$, so that $f'(\mathbf{h}_j \mathbf{g}) = f'(0) + f''(0)\mathbf{h}_j \mathbf{g} + O_P(1/n) = f'(0) + O_P(1/\sqrt{n})$. As a result, the f'-terms in equation (23) are all asymptotically identical and can be removed from the definition of r_k without altering the asymptotics of the variance/mean-squared-error estimators.

When f is linear, $f'(\theta) = f'(0) = 1$, and the r_k in equation (23) are computed as if there were no nonresponse. The same holds true for the he variance/mean-squared-error estimator v_{ST2}. Unfortunately, this f corresponds to an awkward response-probability function: $p_k = 1/(1 + \mathbf{h}_k\gamma)$. Fuller, Loughin, and Baker (1994) made these observations for the case where $\mathbf{h}_k = c_k\mathbf{x}_k$.

The jackknife, v_J, in equation (15) can be computed with these jackknife replicate weights:

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)}/a_k + \Big(\sum_{m\in U}\mathbf{x}_m - \sum_{m\in S} w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m\Big)\Big(\sum_{m\in S} a_{m(\alpha j)}f'(\mathbf{h}_m \mathbf{g})\,\mathbf{h}_m'\mathbf{x}_m\Big)^{-1} c_m a_{k(\alpha j)}f'(\mathbf{h}_k \mathbf{g})\,\mathbf{h}_k'. \qquad (24)$$

Again when $f'(\theta) = f'(0) = 1$, v_J can be computed as if there were no nonresponse.

Like Fuller et al., Folsom and Singh (2000) assumed $\mathbf{h}_k = c_k\mathbf{x}_k$, but they allow the choice of f to vary with k. We leave the straightforward analysis of that extension to the reader. For our purposes, the f in their *generalized exponential (quasi-randomization) model* is

$$f(\mathbf{h}_k\gamma) = \frac{(u-1)(c-\ell)\exp(\mathbf{h}_k\gamma) - (u-c)(1-\ell)}{(u-c) + (c-\ell)\exp(\mathbf{h}_k\gamma)}, \qquad (25)$$

where $\ell \geq 0$, $1 < u \leq \infty$, and $\ell < c \leq u$ are predetermined constants. It is easy to see that when $\mathbf{h}_k\gamma = 0$, $1 + f(\mathbf{h}_k\gamma) = c$; when $\mathbf{h}_k\gamma = -\infty$, $1 + f(\mathbf{h}_k\gamma) = \ell$; and when $\mathbf{h}_k\gamma = \infty$, $1 + f(\mathbf{h}_k\gamma) = u$. Thus, ℓ and u are the lower and upper bounds of $1 + f(\mathbf{h}_k\gamma)$, respectively, while c can be thought of as its center. Folsom and Singh's version of equation (25) replaces c, u, and ℓ with c_k, u_k, and ℓ_k, respectively.

When $c = 1$, $u = \infty$, and $\ell = 0$ in equation (25), $f(\mathbf{h}_k\gamma) = \exp(\mathbf{h}_k\gamma) - 1$, and so $p_k = \exp(-\mathbf{h}_k\gamma)$. That is to say, the log of the probability of response is a linear function of $\mathbf{h}_k$. When $c = 2$, $u = \infty$, and $\ell = 1$, $f(\mathbf{h}_k\gamma) = \exp(\mathbf{h}_k\gamma)$, and $p_k = [1 + \exp(-\mathbf{h}_k\gamma)]^{-1}$, so the probability of response is a logistic function of $\mathbf{h}_k$.

Folsom and Singh pointed out that the treatment of nonresponse through calibration weighting can also be used to adjust for undercoverage. In the

context, the quasi-random phase as sampling occurs conceptually before the actual sample is drawn. The population associated with the sampling frame is assumed to be a Poisson sample from a hypothetical complete population for which the vector T_x must be known. The frame population becomes S', while the hypothetical complete population is U. The probability that element $k \in U$ is in S' is assumed to be modeled correctly by equation (20). If the first (from U to S') and second (from S' to S) phases of sampling are independent, then all the theory developed for using calibration weighting to handle nonresponse carries over to handling undercoverage.

The authors also noted that overcoverage (duplication) or a combination of under and overcoverage can be handled in the same way. The definition of p_k in equation (20) becomes the expected number of times k is in the frame, which can now exceed 1 due to potential duplication.

We have seen that the calibration weights described in this section can produce estimators with good prediction-model-based properties (under equation (2)) when the prediction model is correct (in particular, $E(\epsilon_k \mid \{\mathbf{x}_g, \mathbf{h}_g \mid g \in S\}, \{ I_g \mid g \in U\}) = 0$), and good quasi-randomization properties when the response or coverage model (in equation (20)) is correct. In some sense, one model provides protection against the failure of the other. See Kott (1994).

8. Concluding Remarks

When faced with unit nonresponse, many have attempted to estimate the element probabilities of response, $p_k = 1/[1 + f(\mathbf{h}_k \gamma)]$, directly. This method requires one to have information on $\mathbf{h}_k$ for every element in the sample whether it responded to the survey or not, but $\mathbf{h}_k$ need not have the same dimension as $\mathbf{x}_k$. The direct-adjustment method is generally not available for handling coverage errors.

Fuller (2002) noted that there can be an extra term in the quasi-random mean squared error of

$$t_{y\_GREG} = \sum_S a_k^* y_k + (T_x - \sum_S a_j^* \mathbf{x}_j)(\sum_S c_j a_j^* \mathbf{x}_j')^{-1} \sum_S c_k a_k^* \mathbf{x}_k' y_k,$$

where S is the respondent subsample,
$a_k^* = a_k[1 + f(\mathbf{h}_k \mathbf{g})]$,
and $\mathbf{g}$ is a consistent direct estimator for the quasi-randomization model parameter, γ.

To control the magnitude of the weight adjustment due to nonresponse, Little (1986) recommended that one estimate $\mathbf{g}$ explicitly and then divide the sample into C mutually exclusive cells – often call "poststrata" – based on their fitted $f(\mathbf{h}_k \mathbf{g})$ values. One can then compute the adjusted weight for each element k in cell c as $w_{k\_ADJ} = (\sum_{S'(c)} w_g / \sum_{S(c)} w_g) w_k$, where $S'_{(c)}$ is that part of the original sample in cell c, S(c) is the subsample of S'(c) that respond, and w_k is the sampling weight assigned to element k after sampling but before quasi-random subsampling. This approach assumes that each element in a cell has (roughly) the same probability of response.

Estimating the variance/mean-squared-error of $t_{y\_ADJ} = \sum_S w_{k\_ADJ} y_k$ is beyond the scope of this paper. Whether of not the w_k are calibrated to anything, there is a different calibration after the quasi-random phase, where the w_{k_ADJ} do not allow the estimated number of farms in a cell to change. See Estevao and Särndal (2002) for a discussion of nine different ways to calibrate a two-phase sample.

In the last section we noted that it is possible for components of $\mathbf{h}_k$ in equation (20) to be unknown before response. When such an $\mathbf{h}_k$ is used in calibration, it might no longer to reasonable to assert that the resulting t_{y_CAL} is prediction-model unbiased. This is particularly troublesome when nonresponse is modest compared to the sample size. An intriguing idea is to calibrate in two phases. The first phase adjusts for the difference between T_x and $\sum_{S'} a_k \mathbf{x}_k$, and would not include any components in $\mathbf{h}_k$ unavailable at the time of sampling. The second phase adjusts for the difference between $\sum_{S'} a_k \mathbf{x}_k$ and $\sum_S a_k \mathbf{x}_k$ and would include component variables only available after the respondent subsample is enumerated. A more thorough analysis of this idea must wait for another time.

Let us return to the situation where the response probability in equation (20) is estimated explicitly. An alternative way of incorporating fitted $f(\mathbf{h}_k \mathbf{g})$ values into the estimation presents itself based on methodology developed in the text. Divide the fitted values into P cells, where P is again the dimension of $\mathbf{x}_k$, and let $\mathbf{d}_k$ be a row vector of indicator variables for the P cells. By setting each

$$w_k = a_k[1 + (T_x - \sum_S a_j \mathbf{x}_j)(\sum_S a_j \mathbf{d}_j' \mathbf{x}_j)^{-1} \mathbf{d}_k'],$$

one computes a set of weights for the respondent subsample that, unlike $\{w_{k\_ADJ}\}$ above, satisfies the calibration equation for the respondent sample. Because of the nature of $\mathbf{d}_k$, this linear method returns the same set of calibration weights as fitting $w_k = a_k \exp(\mathbf{d}_k \mathbf{f})$ would – if both produce a set of weights. Note that since calibration weights can be negative with the linear method, it may be able to find a set that the generalized raking method cannot. The linear method effectively scales the a_k-value for every element in the same cell by a fixed amount. Thus, it is unlikely to produce surprisingly small or surprisingly large weights when the dimension of $\mathbf{x}_k$ is small compared to the sample size.

At what point P becomes too large in practice for the sample size – recall P is assumed to stay fixed as

n grows asymptotically large − remains an unanswered component of the broader question of how "best" to create calibration weights. Brewer (private communication) has speculated that P should not exceed $\sqrt{n}$.

One would think that in the absence of nonresponse or coverage errors, a version of Rao's randomization-optimal estimator would be optimal at least in terms of minimizing randomization mean squared error for a given $\mathbf{x}_k$. Recent empirical work by Montanari and Ranalli (2002) show this not always to be the case when the number of strata is large compared to the sample size. Moreover, there are often other considerations: attaining a small model variance for a particular *realized* sample, making sure that no calibration weight is less than 1 (except, perhaps, when adjusting for duplication). A satisfying theory relating $\mathbf{x}_k$, $\mathbf{h}_k$, and f with the size of model variance and/or randomization mean squared error is presently beyond our grasp.

9. References

Asok, C. and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, **71**, 912-918.

Breidt, F.J. and Opsomer, J. D. (2000). Local polynomial regression in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, **74**, 911-915.

Brewer, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1)A 213-233.

Chambers, R.L. (1996). Robust case weighting for multipurpose establishment surveys. *Journal of Official Statistics*, **12**, 3-52.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, *11*, 427-444.

Deville, J-C. and Särndal, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

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

Estevao, V.M., and Särndal, C-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, **18**, 233-255.

Folsom., R.E. and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington DC, 598-603.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. John Wiley & Sons, New York.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, **28**, 5-23.

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

Huang, E.T. and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. In *Proceedings of the Section on Social Statistics*, American Statistical Association, Washington DC, 300-305.

Isaki, C.T. and Fuller, W.A. (1982). Survey design

under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, **24**, 287-296

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, **89**, 693-696.

Kott, P.S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, to appear.

Montanari, G.E. and Ranalli, M.G. (2002). Asymptotically efficient generalised regression estimators. *Journal of Official Statistics*, 18, 577-589.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimations stage. *Journal of Official Statistics*, 25, 1-21.

Särndal, C-E, Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, **76**, 527-537.

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

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology*, **28**, 103-114.