Survey Research Methods (2012)
Vol.6, No.2, pp. 105-111
ISSN 1864-3361
http://www.surveymethods.org

Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine

Phillip S. Kott and Dan Liao
RTI International

Given a randomly drawn sample, calibration weighting can provide double protection against the selection bias resulting from unit nonresponse. This means that if either an assumed linear prediction model or an implied unit selection model holds, the resulting estimator will be asymptotically unbiased in some sense. The functional form of the selection model when using linear calibration adjustment is dubious. We discuss an alternative, nonlinear calibration-weighting procedure and software that can, among other things, implicitly estimate a logistic-response model.

Keywords: logistic response, raking, generalized exponential form, selection model, prediction model, WTADJUST

1 Introduction

Calibration weighting is a method for adjusting the weights in probability-sampling theory by forcing the weighted sum of each of a set of variables to equal specified targets. There are at least two reasons to calibrate survey weights. One is to make estimators unbiased under a linear prediction model. This will often reduce their mean squared errors under probability-sampling theory as well. The other is to adjust for selection bias caused by unit nonresponse or by coverage errors in the frame.

Although it is natural to justify calibration weighting as a nonresponse adjustment tool with a prediction model, it is more common in the survey-sampling literature to argue that a unit's calibration weight adjustment implicitly estimates the inverse of its probability of response. See, for example, Lundströrm and Särndal (1999) or Section 5.1 of Fuller (2009). Unfortunately, the functional form of the selection model in a linear-calibration weighting adjustment allows the implied estimated selection probability to be less than 0 or greater than 1.

The possibility of negative weights is a problem with linear calibration in general, even for surveys without nonresponse. Huang and Fuller (1978) were the first to suggest a method for removing them. Other methods, like that in Park and Fuller (2005), have followed.

We describe here a particular nonlinear calibration-weighting procedure that includes the implicit estimation of the logistic-response model as a special case. This procedure, an extension of the logit (ℓ, u) generalization of raking (Deville and Särndal 1992; Deville et al. 1993) is dubbed here the "generalized exponential form". It is available in the WTADJUST procedure of the computer package SUDAAN®, but has not been directly treated in the refereed literature. For a rigorous treatment of nonlinear calibration in the absence of nonresponse, the reader is directed to Kim and Park (2010). Kott (2006) briefly discussed using nonlinear calibration when there is nonresponse.

We will see that using the generalized exponential form provides *double protection* against nonresponse bias. This means that if either a linear prediction model or an implied selection model holds, then the resulting estimator is asymptotically unbiased in some sense. The term was coined in Kim and Park (2006), but the concept appears to have originated simultaneously in Kott (1994) and Robins et al. (1994) in the treatment of item nonresponse.

In developing what they called the "generalized exponential model" for a stand-alone forerunner to WTADJUST, Folsom and Singh (2000) only discussed near unbiasedness under the combination of probability-sampling theory and a selection model, although they refer to a "superpopulation model", a term usually used to justify prediction modeling in the survey-sampling literature. See, for example, Särndal (1978). The SUDAAN® manual (RTI International 2008) followed suit, calling WTADJUST "model-based" but likewise not addressing the prediction-model properties of its use. We hope to clear up the confusion this has caused for some readers.

Section 2 introduces linear calibration weighting and its relationship to the general regression estimator (GREG). Section 3 discusses double protection against nonresponse bias. Section 4 describes some nonlinear calibration-weighting routines and Section 5 the generalized exponential form. Section 6 addresses unified variance estimation under either the selection or prediction models, while Section 7 explores a small empirical example. Finally, Section 8 offers some comments and concluding remarks.

2 Linear Calibration Weighting

Suppose we have a randomly drawn sample S from a finite population U. In the absence of nonresponse (or measurement error), calibration weighting creates a set of analy-

Contact information: Phillip S. Kott, RTI International, Rockville, MD, e-mail: pkott@rti.org

sis weights, $\{w_k \mid k \in S\}$, not dependent on the survey values of interest that

1. are close to the original design weights, $d_k = 1/\pi_k$, where π_k is the probability of selection for the k^{th} selected sampling unit; and
2. satisfy a set of linear calibration equations, one for each component of $\mathbf{x}_k$, a vector of auxiliary variables:

$$\sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k \tag{1}$$

By "close", we mean that as the sample grows arbitrarily large, the difference between $w_k$and d_k vanishes in probability. For a formal treatment of the assumed asymptotic structure, see Isaki and Fuller (1982) or Kott (2009).

When estimating a population total, $T = \sum_U y_k$, with $\hat{T} = \sum_S w_k y_k$ or a population mean, $\bar{y}_U = T/N$, with $\hat{\bar{y}}_U = \sum_S w_k y_k / \sum_S w_k$, calibration weighting will tend to reduce mean squared error under probability sampling theory when y_k is correlated with the components of $\mathbf{x}_k$. See, for example, Rao (2005).

More formally, $\hat{T}$ is an unbiased estimator for T under the linear prediction model:

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \epsilon_k \tag{2}$$

where $E(\varepsilon_k \mid \mathbf{x}_k) = 0$ whether or not k is in the sample. For our purposes, assuming $E(\varepsilon_k \mid \mathbf{x}_k) = 0$ means that the design is ignorable (given $\mathbf{x}_k$), although that assumption is technically a bit stronger requiring $\varepsilon_k \mid x_k$ to not depend on whether k is in the sample. Since T is itself random under the prediction model, unbiasedness in this context means $E(\hat{T}-T) = 0$.

The simplest way to compute calibration weights is linearly with

$$w_k = d_k \left[1 + \left(\sum_{j\in U} \mathbf{x}_j - \sum_{j\in S} d_j \mathbf{x}_j\right)^T \left(\sum_{j\in S} d_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_k\right] = d_k\left[1 + \mathbf{g}^T \mathbf{x}_k\right],$$

which also produces the generalized regression (GREG) estimator:

$$\hat{T}_{GREG} = \sum_{k\in S} w_k y_k = \sum_{k\in S} d_k y_k + \left(\sum_{j\in U} \mathbf{x}_j - \sum_{j\in S} d_j \mathbf{x}_j\right)^T \left(\sum_{j\in S} d_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \sum_{k\in S} d_k \mathbf{x}_k y_k$$

discussed in Särndal et al. (1989).

As Deville and Särndal (1992) noted when they coined the term, there are many calibration routines where $w_k = d_k f\left(\mathbf{g}^T \mathbf{x}_k\right)$, but $f\left(\mathbf{g}^T \mathbf{x}_k\right) \neq 1 + \mathbf{g}^T \mathbf{x}_k$. These are asymptotically equivalent to the GREG when $f(0) = f'(0) = 1$, $\left|f''(0)\right|$ is bounded, and $\mathbf{g}^T \mathbf{x}_k$ converges to zero as the sample size grows arbitrarily large.

Calibration routines are nonlinear when the *weight-adjustment factor*, $f(.)$, is nonlinear. Despite being nonlinear, these routines produce estimators that are unbiased under the linear prediction model in equation (2).

Although the right-hand side of equation (1) is written as a sum, calibration weighting can be used when the population totals for the auxiliary variables come from an outside source rather than a frame where the individual x-values are known. For our purposes, we will assume that this outside source provides a measure for $\sum_U \mathbf{x}_k$ that has no error.

3 Linear Calibration Weighting for Unit Nonresponse

Most surveys experience unit nonresponse beyond a statistician's control. One is forced to assume, either explicitly or implicitly, some type of model to adjust for the nonresponse. A prediction model (also called an "outcome model") on the survey variable usually assumes the response/nonresponse mechanism, like the sampling design, is ignorable. A selection (or response) model assumes the response mechanism behaves like a phase of Poisson sub-sampling. Double protection means that if *either* the prediction or selection model is specified correctly, the estimator will be nearly unbiased in some sense.

The sample S is replaced by the respondent sample R in defining the GREG,

$$\hat{T}_{GREG} = \sum_{k\in R} w_k y_k = \sum_{k\in R} d_k (1 + \mathbf{g}^T \mathbf{x}_k) y_k,$$

where $\mathbf{g}$ is now

$$\mathbf{g} = \left(\sum_{j\in R} d_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \left(\sum_{j\in U} \mathbf{x}_j - \sum_{j\in R} d_j \mathbf{x}_j\right)$$

or

$$\mathbf{g} = \left(\sum_{j\in R} d_j \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \left(\sum_{j\in S} d_j \mathbf{x}_j - \sum_{j\in R} d_j \mathbf{x}_j\right)$$

depending on whether the respondent sample is calibrated to the population:

$$\sum_{k\in R} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k \tag{3}$$

or to the original sample

$$\sum_{k\in R} w_k \mathbf{x}_k = \sum_{k\in S} d_k \mathbf{x}_k. \tag{4}$$

When calibrating to the population, the estimator $\hat{T}_{GREG}$ is unbiased under the prediction model in equation (2) whether or not k is in the respondent sample (i.e., the response mechanism is ignorable). When calibrating to the sample, it is not hard to see that the estimator is unbiased under a combination of the prediction model and the original sampling design (i.e., probability-sampling theory). We need the latter because when calibrating to the sample, equation (1) need only hold on average across all samples.

Either way, as Fuller et al. (1994) observed the estimator is also nearly unbiased under the quasi-sample design that treats response as a second phase of random sampling as long as each population unit's probability of response if sampled has the form

$$p_k = \frac{1}{1 + \boldsymbol{\gamma}^T \mathbf{x}_k}$$

and $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$. Put another way,

$$\hat{T}_{GREG} = \sum_{k \in R} w_k y_k = \sum_{k \in R} d_k \frac{1}{\hat{p}_k} y_k = \sum_{k \in R} d_k (1 + \mathbf{g}^T \mathbf{x}_k) y_k.$$

Notice that with nonresponse $\mathbf{g}^T \mathbf{x}_k$ does not converges to 0 when calibrating for nonresponse.

4 Nonlinear Calibration

4.1 Raking

Raking is a form of nonlinear calibration in which effectively the calibration weights have the form:

$$w_k = d_k \exp(\mathbf{g}^T \mathbf{x}_k). \tag{5}$$

Traditionally, the components of $\mathbf{x}_k$ are 0/1 indicator variables, and an iterative proportional fitting routine is used to solve the calibration equations (Deming and Stephan 1940). The components do not have to be binary, however, although a component of $\mathbf{x}_k$ should contains all 1's or the equivalent (i.e., there should be a vector $\mathbf{q}$ such that $\mathbf{q}^T \mathbf{x}_k = 1$ for all k). Following Folsom (1991), an iterative process of successive linearizations – Newton's method – can often find a $\mathbf{g}$ that satisfies either the calibration equations in (3) or (4).

Using raking to adjust the weights results in a calibration estimator, $\hat{T} = \sum_S w_k y_k$, with the same unbiasedness properties with respect to the linear prediction model in equation (2) as the GREG. When combined with the original sampling design, however, the quasi-random selection model under which the estimator is nearly unbiased has a slightly more reasonable form than the GREG. It is

$$p_k = \exp(-\boldsymbol{\gamma}^T \mathbf{x}_k),$$

which cannot be less than 0, although it can annoyingly exceed 1.

Raking produced an estimator that is asymptotically equivalent to the GREG when there is no nonresponse or when the population is divided into mutually exclusive groups and each population unit in a group is equally likely to response when sampled (i.e., under a conventional poststratification or reweighting-cell environment). Otherwise, it may not.

4.2 A Logistic-Response Model

Folsom (1991) also proposed using Newton's method to find a $\mathbf{g}$ that forces the w_k to satisfy either equation (3) or (4) such that

$$w_k = d_k \left[1 + \exp\left(-\mathbf{g}^T \mathbf{x}_k\right)\right]. \tag{6}$$

Like with raking, the calibration estimator, $\hat{T} = \sum_S w_k y_k$, has the same unbiasedness properties as the GREG with respect to the linear prediction model in equation (2). Now, however, the estimator is also nearly unbiased under the combination of the original sampling design and the logistic-response model:

$$p_k = \left[1 + \exp\left(-\gamma^T \mathbf{x}_k\right)\right]^{-1} = \frac{\exp\left(\gamma^T \mathbf{x}_k\right)}{1 + \exp\left(\gamma^T \mathbf{x}_k\right)},$$

where $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$. Although not a maximum-likelihood (ML) method, Kim and Riddles (2012) showed empirically that finding the $\mathbf{g}$ satisfying equations (6) and (4) can produce a better estimator for T than estimating $\boldsymbol{\gamma}$ using maximum likelihood. This was likely the result of the calibration-weighted estimator, but not the one using ML, being unbiased under the combination of the original sampling design and a linear prediction model that roughly held in the data they were analyzing.

The weight-adjustment factors in equation (6) are "centered" at 2 in the sense that $f(0) = 2$. By contrast, raking and GREG adjustments are centered at 1.

4.3 Some Generalizations

Observe that the logistic weight-adjustment factor in equation (6) cannot be less than 1, while the raking weight adjustment in equation (5) cannot be less than 0. The GREG weight-adjustment factor, by contrast, can be negative. None of the three weight-adjustment factors have an upper bound.

A useful generalization of raking, called the "logit (ℓ, u)" by Deville et al. (1993) and implemented by them in the SAS program CALMAR, bounds the weight-adjustment factor between ℓ and u:

$$f(\mathbf{g}^T \mathbf{x}_k) = \frac{\ell(u-1) + u(1-\ell)\exp(A\mathbf{g}^T \mathbf{x}_k)}{(u-1) + (1-\ell)\exp(A\mathbf{g}^T \mathbf{x}_k)}, \tag{7}$$

where

$$A = \frac{u - \ell}{(1-\ell)(u-1)},$$

and $\infty \geq u > \ell \geq 0$. This choice of A simplifies finding a derivative for $f(.)$, which is needed for the series of linearizations made when applying Newton's method. We will have need of that derivative in the next section.

Raking is the extreme case of equation (7) where $\ell = 0$ and $u = \infty$. Like raking, however, the logit (ℓ, u) is centered at 1. In fact, when there is no nonresponse, it is asymptotically equivalent to the GREG.

A further extension of the weight-adjustment factor in equation (7) found in Kott (2006) replaces 1 with a centering parameter c:

$$f(\mathbf{g}^T \mathbf{x}_k) = \frac{\ell(u-c) + u(c-\ell)\exp(A\mathbf{g}^T \mathbf{x}_k)}{(u-c) + (c-\ell)\exp(A\mathbf{g}^T \mathbf{x}_k)}, \tag{8}$$

where now

$$A = \frac{u-\ell}{(c-\ell)(u-c)},$$

and $\infty \geq u > c > 1$. Equation (6), the weight-adjustment factor for the logistic-response model, is the special case of equation (8) where $u{=}\infty$, $c = 2$, and $\ell = 1$.

By using equation (8) to adjust for nonresponse, not only can the weight-adjustment factor be centered at a value other than 1, the probabilities of response can be bounded from below by $1/u$ and from above by $1/\ell$.

When adjusting for nonresponse, it seems to make sense to center the weight adjustment at the inverse of the overall response rate as suggested by Folsom and Witt (1994). It turns out the choice of c doesn't matter as long as it is between ℓ and u and $\mathbf{x}_k$ contains an intercept term or the equivalent (i.e., there is a vector $\mathbf{q}$ such that $\mathbf{q}^T\mathbf{x}_k{=}1$ for all k). A little algebra shows that and any choice of c between ℓ and u is equivalent to assuming a response model of the form.

$$p_k = \frac{1+\frac{\exp(\gamma^T\mathbf{x}_k)}{u}}{\ell+\exp(\gamma^T\mathbf{x}_k)},$$

Since $[u(c{-}\ell)/(u{-}c)]\exp(q) = \exp(\log[u(c{-}\ell)/(u{-}c)]{+}q)$, the choice of c only effects the coefficient of the intercept in $\gamma^T\mathbf{x}_k$.

5 The Generalized Exponential Form

Equation (8) is actually a special case of the *generalized exponential form* in Folsom and Singh (2000). Those authors used the word "model" in place of "form", but since their factor can be used when there is no nonresponse and nothing is being modeled, we call it a "form" here.

The generalized exponential form allows each k to have its own u, c and ℓ value:

$$f_k(\mathbf{g}^T\mathbf{x}_k) = \frac{\ell_k(u_k-c_k)+u_k(c_k-\ell_k)\exp(A_k\mathbf{g}^T\mathbf{x}_k)}{(u_k-c_k)+(c_k-\ell_k)\exp(A_k\mathbf{g}^T\mathbf{x}_k)}, \tag{9}$$

where

$$A_k = \frac{u_k-\ell_k}{(c_k-\ell_k)(u_k-c_k)},$$

and $\infty \geq u_k > c_k > \ell_k \geq 0$. Although the centering and bounding parameters can vary across the k, the $\mathbf{g}$ is a constant. To find it using Newton's method, it is helpful to realize

$$f_k'(z) = \frac{(u_k-f_k(z))(f_k(z)-\ell_k)}{(u_k-c_k)(c_k-\ell_k)}, \tag{10}$$

which is both always positive and bounded from above. Since the weight-adjustment factors in equations (6), (7), and (8) are special cases of (9), their derivatives are special cases of (10). The form of the weight-adjustment factor in equation (9) allows the user to bound the weights themselves. For example, setting $\ell_k = 1/d_k$ forces all weights to be no smaller than unity. Similarly, setting $u_k = Q/(d_k y_k)$, when possible (recall u_k must exceed c_k), keeps the weighed survey totals for a key survey variable – the $w_k y_k$ – no greater than Q. Notice, however, that strict unbiasedness under the prediction model is lost with this upper bound because the calibration adjustments depend on the survey values.

When calibrating for consistency with outside sources or for mean squared error reduction in the absence of nonresponse, one can center at 1, like the GREG, and make use of the bounding properties of the weight-adjustment factor in equation (9) to reduce the impact or large weights on mean squared errors. When adjusting for nonresponse, however, it appears more appropriate to employ the special case of the generalized exponential form in equation (8) where the bounding parameters are each constant across k and the centering parameter doesn't matter.

6 Variance Estimation

We have been claiming that $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$ under the combination of the original sample design and the selection model (with mild restrictions on the population and design). We will now sketch the reasoning behind that claim when the unit respondents are calibrated to the original sample assuming that the selection model for unit response implicitly assumed by the parameter settings of the generalized exponential form is correct. In the absence of frame errors or unit nonresponse, $\mathbf{g}$ "estimates" $\mathbf{0}$.

The derivative in equation (10) is always positive and bounded from above. As a result, the mean value theorem tells us that there is a θ_k between $\mathbf{g}^T\mathbf{x}_k$ and $\gamma^T\mathbf{x}_k$ such that $f_k(\mathbf{g}^T\mathbf{x}_k) = f_k(\gamma^T\mathbf{x}_k){+}f_k'(\theta_k)(\mathbf{g}-\gamma)^T\mathbf{x}_k$. Since the calibration equations in (4) hold:

$$\mathbf{g}{-}\gamma = -\left(\sum_{k\in R} d_k f_k'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T\right)^{-1}\left(\sum_{k\in R} d_k f_k(\gamma^T\mathbf{x}_k)\mathbf{x}_k - \sum_{k\in S} d_k\mathbf{x}_k\right),$$

as long as $\sum_R d_k f_k'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T$ is invertible. If in addition, $\sum_R d_k f_k'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T/N$ converges to a positive definite matrix $\mathbf{F}$ and $\sqrt{n}(\sum_R d_k f_k(\gamma^T\mathbf{x}_k)\mathbf{x}_k - \sum_S d_k\mathbf{x}_k)/N$ converges to a bounded vector as the sample size n grows arbitrarily large (the latter because the implicitly assumed response model holds), $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$ such that $\mathbf{g}-\boldsymbol{\gamma} = \mathrm{O}_P(1/\sqrt{n})$. For simplicity, we are assuming that when the sample is multistage, the number of primary sampling units grow in proportion to the sample size.

Since $\sum_R d_k f_k'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T/N$ is within $\mathbf{O}_P(1/\sqrt{n})$ of $\mathbf{F}$ under mild conditions we assume to hold,

$$\begin{aligned}\mathbf{Var}(\mathbf{g}) &\approx \frac{1}{N^2}\mathbf{F}^{-1}\mathbf{Var}\left(\sum_{k\in R} d_k f_k(\gamma^T\mathbf{x}_k)\mathbf{x}_k - \sum_{k\in S} d_k\mathbf{x}_k\right)\mathbf{F}^{-1}\\ &= \frac{1}{N^2}\mathbf{F}^{-1}\mathbf{Var}\left(\sum_{k\in R} d_k\frac{1}{p_k}\mathbf{x}_k - \sum_{k\in S} d_k\mathbf{x}_k\right)\mathbf{F}^{-1}.\end{aligned} \tag{11}$$

Notice that the last expression on the right-hand side of equation (11) has only one measurable source of randomness: $\sum_R d_k p_k^{-1}\mathbf{x}_k$ as an estimator for $\sum_S d_k\mathbf{x}_k$ under the selection model (when calibrating to the population, $\sum_R d_k p_k^{-1}\mathbf{x}_k$ is an estimator for $\sum_U \mathbf{x}_k$ under the original sample and the selection model). To estimate the variance/mean-squared-error estimator for $\mathbf{g}$, useful for determining which variables are significant causes of selection bias if untreated, we can replace the fixed $\mathbf{F}$ by the nearly identical $\sum_R d_k f_k'(\mathbf{g}_k^T\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^T/N$and $f_k(\gamma^T\mathbf{x}_k)$ by $f_k(\mathbf{g}^T\mathbf{x}_k)$.

Developing an estimator for the variance of $\hat{T} = \sum_R w_k y_k$ is simpler when assuming the prediction model than when assuming the selection model. This is because when the prediction model holds the estimator can be rewritten as $\hat{T} = \sum_R w_k\mathbf{x}_k^T\boldsymbol{\beta} + \sum_R w_k\varepsilon_k = \sum_S d_k\mathbf{x}_k^T\boldsymbol{\beta} + \sum_R w_k\varepsilon_k$. The prediction-model variance of $\hat{T}$ is the variance of $\sum_R w_k\varepsilon_k$, while the added variance due to the original sampling design is the design variance of $\sum_S d_k x_k^T\boldsymbol{\beta}$. Any weighted regression estimator $\mathbf{b}$ can be used in variance estimation in place of prediction-model parameter $\boldsymbol{\beta}$ while $y_k - \mathbf{x}_k{}^T\mathbf{b}$ is used in place ϵ_k.

We can do something seemingly analogous in developing a variance/mean-squared-error estimator for $\hat{T}$ under the combination of the original sampling design and the selection model. One important difference is that we focus on a particular form of the "weighted regression estimator": $\mathbf{b} = \left[\sum_R d_k f_k'(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1}\sum_R d_k f_k'(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k y_k$ for a reason that will become clear in the next few paragraphs.

Since we are no longer assuming the prediction model holds, $\mathbf{b}$ does not converge to the model parameter $\boldsymbol{\beta}$. Instead, we assume that under the combination of the original design and the selection model, $\mathbf{b}$ converges to a value $\mathbf{b}^*$ and that $\mathbf{b} - \mathbf{b}^* = \mathbf{O}_P(1/\sqrt{n})$.

The calibration estimator can be expressed as:

$$\hat{T} = \sum_{k\in S} d_k\mathbf{x}_k^T\mathbf{b}^* + \sum_{k\in R} d_k f_k(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{x}_k^T\mathbf{b}^*)$$

$$= \sum_{k\in S} d_k\mathbf{x}_k^T\mathbf{b}^* + \sum_{k\in R} d_k f_k(\gamma^T\mathbf{x}_k)(y_k - \mathbf{x}_k^T\mathbf{b}^*) +$$

$$\sum_{k\in R} d_k f_k'(\theta_k)[(\mathbf{g}-\gamma)^T\mathbf{x}_k](y_k - \mathbf{x}_k^T\mathbf{b}^*)$$

$$= \sum_{k\in S} d_k\mathbf{x}_k^T\mathbf{b}^* + \sum_{k\in R} d_k f_k(\gamma^T\mathbf{x}_k)(y_k - \mathbf{x}_k^T\mathbf{b}^*)$$

$$+ \sum_{k\in R} d_k f_k'(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{x}_k^T\mathbf{b})[(\mathbf{g}-\gamma)^T\mathbf{x}_k] + \mathrm{O}_P(1/n)$$

$$= \sum_{k\in S} d_k\mathbf{x}_k^T\mathbf{b}^* + \sum_{k\in R} d_k p_k^{-1}(y_k - \mathbf{x}_k^T\mathbf{b}^*) + \mathrm{O}_P(1/n).$$

The key step here is that $\mathbf{b}$ has been defined so that ${}_R d_k f_k$' $(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k(y_k - \mathbf{x}_k{}^T\mathbf{b}) = 0$.

The variance/mean squared error of $\hat{T}$ under the original design and the selection model is the equivalent of the variance $\Sigma_S d_k z_k{}^*$, where $z_k{}^* = \mathbf{x}_k^T\mathbf{b}^* + p_k^{-1}(y_k - \mathbf{x}_k^T\mathbf{b}^*)I_k$ and $I_k = 1$ when k is a unit respondent and 0 otherwise. For many designs, this can be estimated by replacing $\mathbf{b}^*$ with $\mathbf{b}$, and estimating the variance of $\Sigma_S d_k z_k$ under the original design as if the $z_k = \mathbf{x}_k{}^T\mathbf{b} + p_k{}^{-1}(y_k - \mathbf{x}_k{}^T\mathbf{b})I_k$ were constants. As Kott (2006) observed, this estimator can also serve as an estimator for the combined variance of $\hat{T}$ under the original design and the prediction model.

7 A Small Empirical Example

We performed a limited simulation study similar to the one in Kim and Park (2010). An artificial finite population of size N= 10,000 was created. Population value were generated from $z_k \sim$ exponential (1), and $y_k|z_k = 2 + z_k + e_k$, where the e_k came from a the standard normal distribution. Samples in each of 10,320 simulations were drawn with size n = 200, and respondents generated using the logistic-response function: $p_k = 1/(1 + .2\exp\{2[\log(z_k) + .5]\})$. Response rates varied from 57 to 81.5%, with an average of roughly 70%.

We created calibration weights in six different ways. We calibrated to the sample three times using equation (4) with $\mathbf{x}_k = (1\ \log(z_k))^T$. In the first, we used linear calibration ($w_k = d_k(1+\mathbf{g}^T\mathbf{x}_k)$), in the second the raking form in equation (5), and in the third the logistic form in equation (6).

We also calibrated to the population three times using equation (3) with $\mathbf{x}_k = (1\ z_k)^T$.

Only the third calibration used the correct form (or the selection model), but the fourth, fifth, and sixth produces an unbiased estimate of the population mean of the y_k under the prediction model since the expectation of $y_k\ |\mathbf{x}_k$ is $\mathbf{x}_k{}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (2\ 1)^T$, for both respondents and nonrespondents.

The results in Table 1 bear this out. We discarded slightly less than 1% of the simulations because calibration was not achieved (model did not converge) using one of the nonlinear methods after 20 iterations. The table displays averages among a random 10,000 of the remaining 10,205 simulations.

Using the last four calibration methods, the relative (empirical) biases of the population y-mean estimates are trivial components of (empirical) relative mean squared errors. Using either of the first two, by contrast, results in a meaningful negative bias. Thus, as expected, if either the selection or prediction model used in the calibration is correct, the resulting estimates are nearly unbiased. If, however, the survey value is not a linear function of the calibration variables in the selection model, it appears important to specify the functional form of the selection model correctly.

The variance estimates described in Section 6 also seem to work adequately for the last four calibration methods (recall that we are really interested in root mean squared errors), although each under-estimates. This is likely caused by using sample residuals (i.e., $y_k - \mathbf{x}_k{}^T\mathbf{b}$) in place of prediction-model errors (ϵ_k) or infinite-population residuals ($y_k - \mathbf{x}_k{}^T\mathbf{b}$*). See Kott (2009) for a further discussion of this problem.

Since our simulations featured with equal probabilities of selection and e_k with constant variances, it should not be surprising that linear calibration to the population on $\mathbf{x}_k = (1\ z_k)^T$ produced the estimated y-means with the smallest mean squared errors overall. Using this method of calibration produced, on average, only 0.4 respondents with negative calibration weights, although 52.8, on average, had calibration weights less than the design weight of 50.

Table 1: Results for estimating $\Sigma^N y_k/N$ *from 10,000 Simulated Simple Random Samples*

Calibration Method	Relative Bias ×100	Relative MSE ×10000	Estimated Variance ×10000
To the sample with $\mathbf{x}_k$=$(1\ \log(z_k))^T$			
Linear	-6.7468	58.351	13.058
Raking	-5.3947	42.593	13.655
Logistic	0.4442	18.504	15.858
To the population with $\mathbf{x}_k$=$(1\ z_k)^T$			
Linear	-0.0383	11.157	10.787
Raking	-0.0147	11.726	10.295
Logistic	-0.0256	15.298	11.519

N=10,000; n= 200; response probability: p_k= 1/(1 + .2exp{2[log(z_k) + .5]}).
$z_k \sim$ exponential (1), $y_k|z_k = 2 + z_k + e_k$, and $e_k \sim$ N(0, 1).

Using the raking method and calibrating to the population could never return negative calibration weights, but 42.4 calibration weights were smaller than 50, on average. When calibrating to the sample on $\mathbf{x}_k = (1\ \log(z_k))^T$, raking had 31.5 respondents with calibration weights less than 50, on average, while linear calibration had 24.7 with calibration weights below 50, 2.0 of which had negative values.

8 Some Comments and Concluding Remarks

8.1 Coverage Errors

Suppose there is undercoverage or duplication in the frame but the population totals for the components of the vector of auxiliary variables $\mathbf{x}$ are (assumed) known. Calibration weighting can be used to adjust for the coverage errors. In the absence of nonresponse, the calibration equations are expressed by (1), while the prediction model remains equation (2).

From a selection modeling viewpoint, finding the $\mathbf{g}$ in equation (8) that satisfies equation (1) can be viewed as implicitly estimating the expected number of times population unit k is on the frame, which has the form:

$$e(\gamma^T \mathbf{x}_k) = \frac{(u-c) + (c-\ell)\exp(A\gamma^T \mathbf{x}_k)}{\ell(u-c) + u(c-\ell)\exp(A\gamma^T \mathbf{x}_k)} \qquad (12)$$

When there is potential frame duplication, the estimated value $e(\mathbf{g}^T\mathbf{x}_k)$ can be greater than 1, and ℓ should accordingly be set at a value less than 1. As with calibration weighting for unit nonresponse, when either the prediction model or selection model in equation (12) is correct, the resulting estimator is unbiased in some sense.

8.2 Weight Trimming

WTADJUST allows the user to trim "extreme" (most often, unusually large) weights before calibration adjustment (e.g., replace d_k in $w_k = d_k f\left(\mathbf{g}^T\mathbf{x}_k\right)$ with a smaller value). As long as the decision about what constitutes an extreme weight is not dependent on the y-value, weighting trimming of this sort does not affect the unbiasedness of an estimator under the prediction model in equation (2) when calibrating the post-trimmed weights using (1) or (3) since the right-hand sides of the calibration equations are unaffected by the trimming. It is more difficult, however, to justify pre-calibration weight trimming under probability sampling theory.

Since probability-sampling theory depends on asymptotic principles, one can mount a probability-sampling defense of weight trimming if it is done rarely. The rule used to determine when a weight needs trimming should be such that the determination tends to disappear as the sample grows arbitrarily large. One such rule would be to trim any d_k that is over a fixed proportion (say 5%) of the population size. A common asymptotic framework in probability sampling theory assumes that the ratio d_k /N approaches 0 as the sample size grows infinitely large. For an establishment survey with a size measure, z_k, one might want to trim any d_k where $d_k z_k/\Sigma_S\, d_j z_j$ or $d_k z_k/\Sigma_U z_j$ is overly large.

8.3 Concluding Remarks

Although we have argued here that using particular parameter settings of the generalized exponential form for the weight adjustment factor can provide double protection against the selection bias due to nonresponse, it is important to remember George Box's maxim: All models are wrong – but some are useful. It is unlikely that all the survey variables of interest follow the same linear prediction model, especially categorical ones. Moreover, even if the selection model were correctly specified by the inverse of the generalized exponential form, what is the likelihood that the bounding parameters have been set correctly? Still, to the extent that either the linear prediction model or the specified selection model approximates reality, calibration weighting as described in the text should, at least, reduce the impact of selection bias due to nonresponse. With this in mind, allowing different setting for the bounding parameters when adjust-

ing for nonresponse is a sensible strategy. For some insight into this issue based on practical experience see Chen et al. (2000).

One problem we did not address is that there may be more than one calibration step. Folsom and Singh (2000), for example, envisioned calibrating the unit respondents first to the full sample to adjust for nonresponse and then to the population to make the estimates consistent with outside sources or, perhaps, adjust for coverage errors in the frame. In addition, there can be nonresponse at multiple points in some surveys, at a screening pre-survey, at the household for screener respondents, and at the individual for household respondents. Each point can require its own calibration step. Vaish et al. (2000) describe a clever way to linearize the variance under the combination of the original design and the various selection models. Using some form of replication may be more advisable in this context, however.

References

Chen, P., Penne, M. A., & Singh, A. C. (2000). *Experience with the Generalized Exponential Model for Weight Calibration for the National Household Survey on Drug Abuse.* Proceedings of the American Statistical Association, Survey Research Methods Section, 604-609.

Deming, W. E., & 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., & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, *88*, 1013-1020.

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

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

Folsom, R. E., & Witt, M. B. (1994). *Testing a New Attrition Nonresponse Adjustment Method for SIPP.* Proceedings of the American Statistical Association, Survey Research Methods Section, 428-433.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: Wiley.

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

Huang, E. T., & Fuller, W. A. (1978). *Nonnegative Regression Estimation for Sample Survey Data.* Proceedings of the American Statistical Association, the Social Statistics Section, 300-305.

Isaki, C. T., & Fuller, W. A. (1982). Survey Design under the Regression Super-population Model. *Journal of the American Statistical Association*, *77*, 89-96.

Kim, J. K., & Park, H. (2006). Imputation Using Response Probability. *Canadian Journal of Statistics*, *34*, 1-12.

Kim, J. K., & Park, M. (2010). Calibration Weighting in Survey Sampling. *International Statistical Review*, *78*, 21-39.

Kim, J. K., & Riddles, M. K. (2012). Some Theory for Propensity Score Adjusted Estimators. *Survey Methodology*. (forthcoming)

Kott, P. S. (1994). A Note on Handling Nonresponse in Surveys. *Journal of the American Statistical Association*, *89*, 693-696.

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

Kott, P. S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics 29b: Sample surveys: Inference and analysis.* New York: Elsevier.

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

Park, M., & Fuller, W. A. (2005). Towards Nonnegative Regression Weights for Survey Samples. *Survey Methodology*, *31*, 85-93.

Rao, J. N. K. (2005). Interplay Between Sample Survey Theory and Practice: An Appraisal. *Survey Methodology*, *31*, 117-138.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, *89*, 846-866.

RTI International. (2008). *SUDAAN Language Manual, Release 10.0.* Research Triangle Park, NC: RTI International.

Särndal, C.-E. (1978). Design-based and Model-based Inference in Survey Sampling (with discussion). *Scandinavian Journal of Statistics*, 27-52.

Särndal, C.-E., Swensson, B., & Wretman, J. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator. *Biometrika*, *76*, 527-537.

Vaish, A. K., Gordek, H., & Singh, A. C. (2000). *Variance Estimation for Weight Calibration via the Generalized Exponential Model with Application to the National Household Survey on Drug Abuse.* Proceedings of the American Statistical Association, Survey Research Methods Section, 616-621.