# Exploring Several Uses of Instrumental-Variable Calibration Weighting

Phillip S. Kott
pkott@rti.org

# Outline

- What is calibration weighting?
- Linear calibration
- A useful nonlinear form and its generalization
- Instrumental Variables (IV's) when there is unit nonresponse
- Using IV's for (nearly) pseudo-optimal calibration
- Using IV's to ease replication
- SUDAAN 11 with some examples
- Concluding Remarks

# What is calibration weighting?

In the absence of nonresponse, calibration is a weight adjustment method that creates a set of weights, $\{w_k\}$, that

1. Are asymptotically close to the original design weights: $d_k = 1/\pi_k$, so that resulting estimates are *nearly unbiased* under probability-sampling theory.

2. Satisfy a set of calibration equations (one for each components of $\mathbf{z}_k$):

$$\sum_S w_k \mathbf{z}_k = \sum_U \mathbf{z}_k$$

When estimating $T = \sum_U y_k$ with $t = \sum_S w_k y_k$

or

$\bar{y}_U = T/N$ with $\sum_S w_k y_k / \sum_S w_k$,

**calibration weighting** will tend to reduce mean squared error when $y_k$ is correlated with components of $\mathbf{z}_k$

(but a real survey has many $y_k$'s).

One way to compute calibration weights is linearly:

$$w_k = d_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} d_k \mathbf{z}_k$$

since $\sum_S w_k \mathbf{z}_k^T = \sum_S d_k \mathbf{z}_k^T +$

$$\left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} \sum_S d_k \mathbf{z}_k \mathbf{z}_k^T$$

Note that if we replace $\mathbf{z}_k^T$ by $y_k$, we have a GREG

Looking at it another way:

$$w_k = d_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j{}^T\right)^{-1} d_k \mathbf{z}_k$$

$$= d_k \left[1 + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j{}^T\right)^{-1} \mathbf{z}_k\right]$$

$$= d_k \left[1 + \mathbf{g}^T \mathbf{z}_k\right]$$

# How Linear Calibration Handles Unit Nonresponse

The vector $\mathbf{g}$ becomes

$$\mathbf{g} = \left(\sum_U \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T \left(\sum_R d_j \mathbf{z}_j \mathbf{z}_j{}^T\right)^{-1} \quad \text{or}$$

$$= \left(\sum_S d_j \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T \left(\sum_R d_j \mathbf{z}_j \mathbf{z}_j{}^T\right)^{-1},$$

depending on whether the respondent sample $R$ is calibrated to the population $\left(\sum_U \mathbf{z}_j\right)$ or to the original sample $\left(\sum_S d_j \mathbf{z}_j\right)$.

Either way, the estimate is nearly unbiased under a quasi-sample-design where unit *k* has probability of response:

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

and **g** is a consistent estimator for **γ.**

The estimated probability of response is not bounded.

It can be less than 0 or greater than 1.

# A Useful Nonlinear Form for Calibration Weighting

Find a vector **g** (through repeated linearized approximations) such that

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{z}_k)\mathbf{z}_k = \sum_U \mathbf{z}_k \quad \text{or} \quad \sum_S d_k \mathbf{z}_k ,$$

where the *weight-adjustment function* is

$$\alpha(\mathbf{g}^T \mathbf{z}_k) = \frac{\ell(u-c) + u(c-\ell)\exp(A\mathbf{g}^T \mathbf{z}_k)}{(u-c) + (c-\ell)\exp(A\mathbf{g}^T \mathbf{z}_k)},$$

and $A = (u-\ell)/[(u-c)(c-\ell)]$.

This adjustment is centered at $c$ (i.e., $\alpha(0) = c$),

with a lower bound $\ell \geq 0$,

and an upper bound $u > c > \ell$.

The user sets these three parameters.

This is a generalization of both raking, where

$\ell = 0,\ c = 1,\ u = \infty,$

and of the implicit estimation of a

logistic-regression **selection** (response) model, where

$\ell = 1,\ c = 2,\ u = \infty.$

# The General Exponential Model

Allows separate weight-adjustment functions for each $k$:

$$\alpha_k(\mathbf{g}^T\mathbf{z}_k) = \frac{\ell_k(u_k - c_k) + u_k(c_k - \ell_k)\exp(A_k\mathbf{g}^T\mathbf{z}_k)}{(u_k - c_k) + (c_k - \ell_k)\exp(A_k\mathbf{g}^T\mathbf{z}_k)}$$

but with a common $\mathbf{g}$ chosen to satisfy one of the two versions of the calibration equation – calibrating to the population or to the original sample.

# Instrumental Variables and Nonresponse

We can assume a selection model:

$$p_k = \left[\alpha(\mathbf{g}^T\mathbf{x}_k)\right]^{-1} = \frac{(u-c)+(c-\ell)\exp(A\mathbf{g}^T\mathbf{x}_k)}{\ell(u-c)+u(c-\ell)\exp(A\mathbf{g}^T\mathbf{x}_k)},$$

where *some* components of the **model** vector $\mathbf{x}_k$

do not coincide with the **calibration** vector $\mathbf{z}_k$.

That is,

$$\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha(\mathbf{g}^T\mathbf{x}_k)\mathbf{z}_k = \sum_U \mathbf{z}_k \text{ or } \sum_S d_k\mathbf{z}_k .$$

# Instrumental Variables

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{x}_k)\mathbf{z}_k = \sum_U \mathbf{z}_k \text{ or } \sum_S d_k \mathbf{z}_k ,$$

Components of $\mathbf{x}_k$ that are not components of $\mathbf{z}_k$ are called **instrumental variables.**

The name derives from the linear-calibration form, where

$$\sum_S w_k y_k = \sum_S d_k y_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{x}_j \mathbf{z}_j^T\right)^{-1} \sum_S d_k \mathbf{x}_k y_k$$

$$\downarrow$$

$$\mathbf{b}_{IV}$$

In the linear prediction model: $E(y_k|\mathbf{z}_k, \mathbf{x}_k) = \mathbf{z}_k\boldsymbol{\beta}$.

In that context, the components of $\mathbf{z}_k$ are the model variables.

(Instrumental variables not in $\mathbf{z}_k$ are often assumed to be independent of the model error while the component they replace are not.)

In establishment surveys, it often makes sense to calibrate to a size variable – call it $q_k$ – because the main survey variable is nearly linear in the size variable.

*But* response is better modeled as a logistic function of the ***log of the size variable***, so that a one percent increase in the size variable results in a $c$ percent change in the odds of response.

Thus, $\log(q_k)$ is an instrument used in place of $q_k$.

Deville (2000) noted that it is possible for a selection model variable to be known only for respondents. That is, for *nonresponse to not be missing at random.*

# Nearly Pseudo-Optimal Calibration

In the absence of nonresponse and coverage errors, the **pseudo-optimal** calibration estimator calibrates linearly on $\mathbf{z}_k$ but sets $\mathbf{x}_k = (d_k - 1)\mathbf{z}_k$.

This version of calibration minimizes asymptotic mean squared error under some designs (e.g., Poisson sampling). Hence, the the calibration is called *pseudo-optimal* generally.

With the general exponential form centered at 1, one can obtain the asymptotic properties of pseudo-optimal calibration while reasonably bounding the weights; for example, with $1 \leq w_k \leq u_k$. The resulting calibration is *nearly pseudo-optimal*.

# Easier Replication Weights

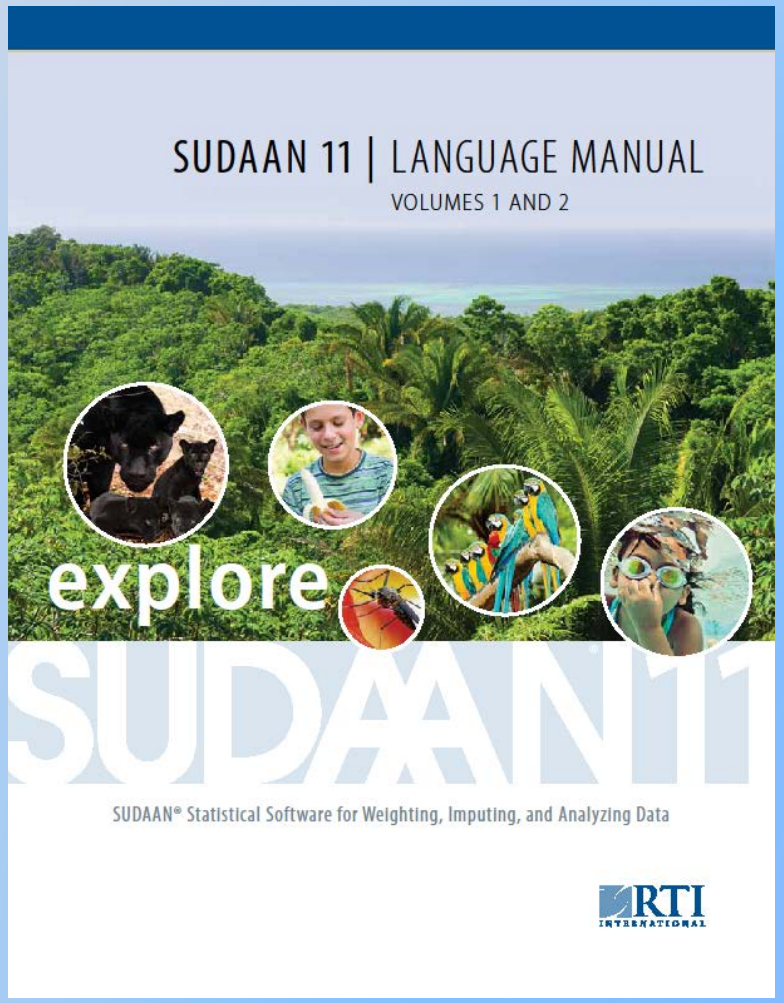
Bounds can be a problem with replicates when using the general exponential form for calibration.

An alternative way to create them is to calibrate centered at 1 ***without bounds***, beginning with $\alpha_k \times d_{k(r)}$ (the pre-adjusted replicate weight), and setting $\mathbf{x}_k = (\phi_k/\alpha_k)\mathbf{z}_k$, where

$$\phi_k = \frac{(\alpha_k - l_k)(u_k - \alpha_k)}{(c_k - l_k)(u_k - c_k)}$$ is the derivate of $\alpha_k(.)$.

(When calibrating for variance reduction only: $\phi_k$ , $\alpha_k \approx 1$.)

Instrumental variable calibration with the general exponential "model" is coming in SUDAAN 11 (PROC WTADJX)



# SUDAAN 11

- Also coming is linearization-based variance estimation when there is one round of calibration or logistic reweighting.

Some effort will be needed when WTADJX is applied for a use other than nonresponse adjustment.

Think of a weight-adjustment function $\alpha_k(\mathbf{g}^T\mathbf{x}_k)$ as a weight *model*.

Instrumental variables = model variables even when there is no selection model.

# An Example Without Nonresponse

A *stsrs* of 364 fictional hospital emergency departments.

A size measure is available on the frame.

**Raking ratio using a size measure** (Hidiroglou and Patak, JOS 2006) by region (four), public/private, and metro/nonmetro:

*Model variables* are public, metro, and region dummies

*Calibration variables* are public × size, metro × size, region × size

A **pseudo-optimal alternative** replaces each model variable with

*variable* × size × (weight – 1)

# *Results (CVs)*

| *Region* | *Unadjusted* | *Raking* | *Nearly Pseudo-optimal* |
|---|---|---|---|
| All | 6.48 | 2.14 | 1.91 |
| 1 | 5.67 | 3.33 | 3.27 |
| 2 | 13.97 | 3.44 | 2.02 |
| 3 | 7.55 | 3.23 | 3.22 |
| 4 | 14.58 | 5.77 | 5.69 |

*survey variable* = annual drug-related visits
*frame size variable* = annual visits of any type in a previous year

# An Example With Nonresponse

Same data set but with nonresponse generated as a logistic function of the *log* of the survey variable (roughly 45 % response).

Assuming first that response is a function of the log of the *frame variable*, we can use SUDAAN to estimate the survey-variable total

Using RLOGIST CV = 7.33

Using WTADJUST CV = 8.30

Using WTADJX

calibrating to the frame variable in the original sample CV = 6.39

calibrating to the frame variable in the population CV = 3.40

We can also use SUDAAN 11 to test whether there is a significant difference between estimates derived under different assumed response models

In this case, the estimated bias (roughly 1.2%) from incorrectly assuming response is a logistic function of the log of the *frame* variable rather than the log of the *survey* variable is significant at the .08 level.

Even when we don't know the true response model, the test – duplicating each record, assigning the first version to a domain governed by one assumed response model and the second to a domain governed by a different assumed model *while keeping both in the same PSU* – can be used to determine whether different response models lead to significantly different estimates.

# Concluding Remarks

Although calibrating to the population is more efficient than calibrating to the full sample, it is better to calibrate in two steps.

That allows one to use nearly pseudo-optimal calibration in the second step and make up for any inefficiency from instrumental-variable calibration.

Empirical research on using instrumental-variables to aid in replication is underway.