Exploring Some Uses for Instrumental-Variable Calibration Weighting

Phillip S. Kott
RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852, U.S.A

Abstract
The new WTADJX procedure in SUDAAN 11 does instrumental-variable calibration weighting using a flexible nonlinear weight-adjustment function. We will review the theory behind this procedure and explore some of its uses.

Key Words: Response model; Generalized exponential form; Standard error; Nearly quasi-optimal calibration; WTADJX.

1. Introduction

Brewer (2000) proposed using instrumental variables in a calibration-weighted estimator for a finite-population total as a way of integrating the prediction form of model-based sampling theory with randomization consistency. We will explore more practical uses of instrumental-variable calibration such as, 1, adjusting for nonresponse when the variables governing the response/nonresponse mechanism are not always the same as the calibration variables, and, 2, creating nearly optimal weights under probability sampling theory that never fall below unity and (if desired) are bounded from above.

Section 2 briefly reviews calibration weighting and the generalized exponential form of Folsom and Singh (2000). Section 3 discusses the hows and whys of instrumental-variable calibration as implemented with the new WTADJX procedure in SUDAAN 11 (RTI 2012). Two examples flesh this out in Section 4. Section 5 offers some concluding remarks.

2. Calibration Weighting

2.1 Linear calibration weighting

In the absence of nonresponse, calibration is a weight-adjustment method that creates a set of weights, $\{w_k\}$ with two important properties.

1. They are asymptotically close to the original design weights: $d_k = 1/\pi_k$ (i.e., as the sample size grows arbitrarily large, w_k converges to d_k) and therefore *nearly* unbiased under probability-sampling theory.

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

$$\sum\nolimits_S w_k \mathbf{z}_k = \sum\nolimits_U \mathbf{z}_k \quad (1)$$

When estimating a total $T = \sum_U y_k$ with $t = \sum_S w_k y_k$ or a 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 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 with the following formula:

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

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

Observe that as the sample size grows arbitrarily large, $\mathbf{g}^T\mathbf{z}_k$ (which means $\mathbf{g}$) converges to $\mathbf{0}$.

This is the weighting scheme implied by the *generalized regression* (GREG) estimator since

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

$$= \sum_S d_k y_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j \right)^T \mathbf{b},$$

where $\mathbf{b} = (\sum_S d_j \mathbf{z}_j \mathbf{z}_j^T)^{-1} \sum_S d_k \mathbf{z}_k y_k$ is a survey-weighted estimated linear-regression coefficient.

Linear calibration weighting can be easily adapted to handle unit nonresponse by simply replacing the sample S with respondent sample R and redefining the GREG estimator and $\mathbf{g}$ as:

$$t_{GREG} = \sum_R w_k y_k = \sum_R d_k \left(1 + \mathbf{g}^T \mathbf{z}_k \right) y_k$$

where

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

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

depending on whether the respondent sample is calibrated to the population $(\sum_U \mathbf{z}_j)$ or to the original sample $(\sum_S d_j \mathbf{z}_j)$. Either way, the estimate is also nearly unbiased under the quasi-sample-design that treats response as a second phase of random sampling as long as each unit's probability of response has the form:

$$p_k = \frac{1}{1 + \boldsymbol{\gamma}^T \mathbf{z}_k}, \qquad (2)$$

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

$$t_{GREG} = \sum_R w_k \mathbf{z}_k = \sum_R d_k \hat{p}_k^{-1} \mathbf{z}_k .$$

Notice that with nonresponse neither

$$(\sum_U \mathbf{z}_j - \sum_R d_j \mathbf{z}_j)^T \text{ nor } (\sum_S d_j \mathbf{z}_j - \sum_R d_j \mathbf{z}_j)^T$$

converges to $\mathbf{0}^T$, and so neither does $\mathbf{g}^T$. This, at the time surprising, use of calibration weighting was proposed by Fuller *et al.* (1994).

2.2 **Nonlinear calibration weighting**

The problem with the probability-of-response function in equation (2) is that it can fall below unity and even be negative. A useful nonlinear form of calibration weighting finds a **g** (through repeated linearization, i.e., Newton's method) such that

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{z}_k) \mathbf{z}_k = \sum_U \mathbf{z}_k \text{ or}$$
$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{z}_k) \mathbf{z}_k = \sum_S w_k \mathbf{z}_k , \tag{3}$$

where $$\alpha(\mathbf{g}^T \mathbf{z}_k) = \frac{\ell(u-c) + u(c-\ell)\exp(A\mathbf{g}^T \mathbf{z}_k)}{(u-c) + (c-\ell)\exp(A\mathbf{g}^T \mathbf{z}_k)} , \tag{4}$$

and $A = (u-\ell)/[(u-c)(c-\ell)]$. The inclusion of A in equation (4) makes taking the derivative of $\alpha(\mathbf{g}^T \mathbf{z}_k)$ easier.

The *weight adjustment* $\alpha(\mathbf{g}^T \mathbf{z}_k)$ is centered at c (i.e., α(0) = 1) with a lower bound $\ell \geq 0$ and an upper bound $u > c > \ell$, which can be infinite. The user sets these *centering* and *bounding* parameters. Equation (4) is a generalization of both raking, where $\ell = 0$, $c = 1$, $u = \infty$ (and the components of $\mathbf{z}_k$ are binary) and the implicit estimation of a logistic-regression response model, where $\ell = 1$, $c = 2$, $u = \infty$.

When c = 1, equation (4) is the generalized-raking adjustment introduced by Deville and Särndal (1992) so that the range of $\alpha(\mathbf{g}^T \mathbf{z}_k)$ could be bounded (and the components of $\mathbf{z}_k$ continuous). Centering at 1 was a requirement of calibration weighting in that landmark paper (α'(0) = 1 was required as well), but setting c > 1 with $\ell = 1$ is more sensible when adjusting for unit nonresponse.

Folsom and Singh (2000) proposed using the following generalized exponential form:

$$\alpha_k(\mathbf{g}^T \mathbf{z}_k) = \frac{\ell_k (u_k - c_k) + u_k (c_k - \ell_k)\exp(A_k \mathbf{g}^T \mathbf{z}_k)}{(u_k - c_k) + (c_k - \ell_k)\exp(A_k \mathbf{g}^T \mathbf{z}_k)} , \tag{5}$$

which allows separate weights functions for each k but finds a common **g** chosen to satisfy one of the two versions of the calibration equation (the population or original-sample version). This form of calibration weighting has been incorporated into the SUDAAN procedure WTADJUST (RTI, 2008). See Kott and Liao (2012) for a more rigorous treatment of this version of nonlinear calibration weighting. Kott (2009) provides good background on calibration weighting *per se*.

Although WTADJUST allows $\alpha_k(\mathbf{g}^T\mathbf{z}_k)$ to be *k*-specific, when adjusting for nonresponse (or coverage), it is sensible to select a single value for the c_k and a very limited number of ℓ_k and u_k values. When each of the three has a single value, it is not hard to see that the choice of c become irrelevant (again, see Kott and Liao, 2012).

3. Instrumental Variables

3.1 Nonresponse

Now suppose unit response followed a model of the form:

$$p_k = \left[\alpha_k(\boldsymbol{\gamma}^T\mathbf{x}_k)\right]^{-1} = \frac{(u_k - c_k) + (c_k - \ell_k)\exp(A_k\boldsymbol{\gamma}^T\mathbf{x}_k)}{\ell_k(u_k - c_k) + u_k(c_k - \ell_k)\exp(A_k\boldsymbol{\gamma}^T\mathbf{x}_k)}, \quad (6)$$

where some components of the *model* vector $\mathbf{x}_k$ need not coincide with the components on the *benchmark* $\mathbf{z}$-vector. In other words, replace equation (2) by

$$\begin{aligned} &\textstyle\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha_k(\mathbf{g}^T\mathbf{x}_k)\mathbf{z}_k = \sum_U \mathbf{z}_k \quad \text{or} \\ &\textstyle\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha_k(\mathbf{g}^T\mathbf{x}_k)\mathbf{z}_k = \sum_S w_k\mathbf{z}_k, \end{aligned} \quad (7)$$

such that $\mathbf{g}$ again estimates $\boldsymbol{\gamma}$.

Mathematically, finding a $\mathbf{g}$ that satisfies either the first or second line of (7) can often be done as long as the number of calibration (benchmark) variables in $\mathbf{z}_k$ is at least as great of the number of model variables in $\mathbf{x}_k$. A routine to do that is available in SUDAAN 11 (RTI 2012): WTADJX. It applies most simply when the numbers of model and calibration variables coincide so that one of the two sets of calibration equations in (7) holds. Otherwise, there are more unknowns than equations, and the vector equations in (7) cannot hold exactly. See Chang and Kott (2008) for a discussion of minimizing the difference between, say, $\sum_R d_k\alpha_k(\mathbf{g}^T\mathbf{x}_k)\mathbf{z}_k$ and $\sum_U \mathbf{z}_k$ as a means for estimating $\boldsymbol{\gamma}$**.**

The 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

$$\begin{aligned} \textstyle\sum_S w_k y_k &= \textstyle\sum_S d_k y_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j\mathbf{z}_j\right)^T \left(\sum_S d_j\mathbf{x}_j\mathbf{z}_j^T\right)^{-1} \sum_S d_k\mathbf{x}_k y_k \\ &= \textstyle\sum_S d_k y_k + \left(\sum_U \mathbf{z}_j - \sum_S d_j\mathbf{z}_j\right)^T \mathbf{b}_{IV} \end{aligned}$$

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. (In econometrics, instrumental variables not in $\mathbf{z}_k$ are often assumed to be independent of the model error while the components 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. Response, by contrast, may be 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.*

Better known is that Brewer (2000) proposed using instrumental-variable calibration as a way of integrating the prediction form of model-based sampling theory with randomization consistency. We will not follow up that idea here.

3.2. Nearly pseudo-optimal calibration

Instrumental-variable calibration can be profitably used in the absence of nonresponse and coverage errors. A linear estimator often better (i.e., more efficient) than the GREG also calibrates on $\mathbf{z}_k$ but sets $\mathbf{x}_k = (d_k - 1)\mathbf{z}_k$. This produces the nearly unbiased linear estimator with the smallest asymptotic mean squared error under Poisson sampling and similarly under stratified simple random sampling with large stratum samples sizes. As a result, it has been called the "optimal estimator" under Poisson sampling (Rao, 1994) and the "pseudo-optimal estimator" more broadly (Banker, 2002).

With WTADJX centered at 1, we can bound the weights and retain the asymptotic properties of the optimal estimator by setting $\mathbf{x}_k = (d_k - 1)\mathbf{z}_k$. In particular, when $d_k > 1$, we can set $\ell_k = 1/ d_k$ to assure that all w_k are at least unity. If some $d_k = 1$, we can simply set $w_k = 1$ and remove *k* from *U* and *S* before applying equation (4) See Kott (2011a). Alternatively, we can simply set ℓ_k at any value less than 1 for elements with $d_k = 1$ since $\mathbf{x}_k$ will be **0** forcing w_k to be 1 as well.

We have some freedom in setting the u_k as long as they are each greater than 1 and the calibration equation ($\sum_S w_k \mathbf{z}_k = \sum_U \mathbf{z}_k$) can be satisfied. Sometimes, rather than bounding the weight adjustment, a user may want to bound the weight itself by creating an upper bound of the form $u_k = U/d_k$. Often with establishment surveys, it is desirable to set an upper bound in of the the form $u_k = \mathrm{U}/(d_k q_k)$ so that $w_k q_k$ is bounded.

4. Two Examples

As has been noted, the WTADJX procedure in SUDAAN 11 can perform instrumental-variable calibration. SUDAAN 11 is also be able to compute (asymptotic) standard errors properly for means, totals, and ratios computed with weights adjusted by one round of WTADJX or WTADJUST calibration (Witt, 2010). When the adjustment is for nonresponse (or coverage error), this assumes that the underlying selection model has been specified correctly, that is, the model in equation (6) holds and that response is independent across primary sampling units. When the logistic response model is correct. SUDAAN will also compute standard errors properly when the LOGISTIC procedure (RLOGIST in the SAS-callable version of SUDAAN used here) is used to adjust the weights.

4.1 An example with no nonresponse or coverage error

We created a stratified simple random sample of 364 fictional hospital emergency departments using the public-use data set of the Drug Abuse Warning Network (US Department of Health and Human Services, 2011) as a starting point. The sample can be found in *SUDAAN 11 Examples* (WTADJX examples) on the SUDAAN website (http://www.rti.org/sudaan/). Much of the SAS-callable SUDAAN code discussed in this Section can be found there as well.

Each hospital on the frame has attached to it a size variable – the number of emergency-department visits in a previous year, which we call "frame visits." There are also indicators on the frame of each hospital's census region, whether it is publicly owned, and whether it is in a metropolitan area. Our goal is the estimate the total number of *drug-related* emergency-department visits in the survey year both across the US and within each census region.

In addition to computing the estimates directly with their probability weights, we "raked" the weights − using WTADJUST with a center of 1, a lower bound of 0, and no upper bound − so that the following calibration-weighted totals equaled the corresponding frame counts: the number of hospitals in each region, the number of public hospitals, and the number of hospitals in a metropolitan area. That is to say, the vector $\mathbf{z}_k$ had six components, four regional indicator dummies, an indicator dummy for public ownership, and an indicator dummy for a metropolitan location.

As can be seen in Table 1, raking did not improve the coefficients of variation (CVs) in any of the regions (computed using SUDAAN 11 as are all the estimates in this section). If anything, the CVs became slightly higher. That is because a hospital's annual number of drug-related emergency-department visits is not nearly a linear function of its region, ownership status, and urbanicity.

A variant of raking for establishment surveys introduced by Hidiroglou and Patak (2006) is more applicable in this setting. *Size raking* calibrates the weights so that the weighted-total of the size variable (q_k) within each region equals the actual number on the frame, with analogous equalities holding for public and metropolitan hospitals. This variant on raking should decrease the standard errors of estimates for drug-related emergency-department visits as the US and regional levels if these survey variables are roughly linear functions of the calibration variables.

Size raking was done in WTADJX by letting the region, public, and metropolitan indicator dummies remain the MODEL variables (with /NOINT since there was no intercept), while each of those indicator dummies times frame visits made up the calibration variables or CALVARS. Here, the "model" refers to the *weight-adjustment model*, $\alpha(\mathbf{g}^T\mathbf{x}_k) = \exp(\mathbf{g}^T\mathbf{x}_k)$, used in WTADJX, where $\mathbf{x}_k$ is the vector of the six indicator dummies, while the vector of calibration variables becomes $\mathbf{z}_k = q_k\mathbf{x}_k$. There is no response (or coverage) model.

Employing size raking decreased CVs noticeably. Better still, as can be see in Table 1, were two variants of nearly-quasi-optimal (NQO) calibration weighting. In one, the same CALVARS were used as in size raking but the MODEL variables were these calibration variables times $d_k – 1$. In the other, an intercept was added. Mathematically, the vector of calibration variables was $(1, \mathbf{z}_k^T)^T$ for the latter version with $\mathbf{z}_k$ defined as in size raking, while the vector of model variables was $(d_k - 1)(1, \mathbf{z}_k^T)^T$. (Ironically, in running WTADJX, /NOINT is still used since $d_k - 1$ appears in the MODEL statement in place of an intercept).

Table 1: Comparing Direct Estimation to Raking and Nearly Quasi-Optimal Calibration

REGION	*Direct*	*Raking*	*Size Raking*	*NQO*	*NQO Intercept*
Estimates					
All	5376256	5371840	5526307	5519244	5531364
East	732957	732749	785407	787582	787026
South	1750451	1746077	1836788	1832655	1833783
Midwest	1369023	1369140	1425517	1426593	1433667
West	1523825	1523874	1478595	1472415	1476887
Coefficients of Variation					
All	6.47	6.48	2.16	1.91	1.87
East	5.67	5.71	3.32	3.27	3.28
South	13.92	13.94	3.49	2.02	1.95
Midwest	7.55	7.55	3.23	3.22	3.26
West	14.58	14.58	5.77	5.69	5.61

4.2 An example with nonresponse

We then used the same data set as in the previous example, but generated unit nonresponse as a logistic function of the *log* of drug-related emergency-department visits. Assuming first that response was a function of the log of the *frame* visits, we employed SUDAAN to estimate survey-variable totals applying first RLOGIST and WTADJUST. We then applied WTADJX, again letting the log of frame visits be the MODEL variable, but now frame visits became the calibration variable (in CALVARS).

These following estimated CVs were computed

Using RLOGIST	CV = 7.33
Using WTADJUST	CV = 8.30
Using WTADJX	
calibrating to the frame visits in the original sample	CV = 6.39
calibrating to the frame visits in the population	CV = 3.40

It may come as a bit of a surprise that adjusting for nonresponse using RLOGIST was estimated to be more efficient than adjusting with WTADJUST. Given, the nature of the data, however, it should be no surprise that using WTADJX and calibrating on frame visits rather than the log of those visits appeared more efficient than using either RLOGIST or WTADJUST even though the same variable (log of frame visits) was used to model response by all three. Moreover, calibrating to frame totals rather than full-sample totals increased the estimated efficiency even more.

WTADJX can be also used 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 was a logistic function of the log of the *frame* variable rather than the log of the *survey* variable was significant at the .08 level.

It would be a mistake to conclude that bias is not an issue here because it statistical significance did not each the magic .05 level. When testing for possible bias we need to be concerned with Type 2 error (failing to recognize a bias when it exists) than Type 1 error (finding a bias when none exists) here. As a result, statistical significance at the .08 level should be viewed as problematic.

Even when we don't know the true response model, the test we used − 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,* treating the sample as if it were drawn with replacement, and then testing the difference between domain estimates − can be applied to determine whether different response models lead to significantly different estimates. A test like this was proposed by Fuller (1984) for determining whether sampling weights matter in a linear regression.

5. 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 to adjust for nonresponse.

Kott (2011b) points out that instrumental-variable calibration can aid in replication when a bounded version of WTADJUST or WTADJX calibration is used. Empirical research on this use of WTADJX is underway.

References

Bankier, M. (2002). Regression Estimators for the 2001 Canadian Census, presented at the International Conference in Recent Advances in Survey Sampling.

Chang, T. and P. S. Kott (2008). Using Calibration Weighting to adjust for Nonresponse Under a Plausible Model, *Biometrika*, 95, 557-571.

Deville, J.-C. (2000). Generalized Calibration and Application to Weighting for Non-response, *Compstat: Proceedings in Computational Statistics; 14th Symposium Held in Utrecht*, *The Netherlands* (available online).

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

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

Fuller, W.A. (1984). Least Squares and Related Analyses for Complex Survey Designs, *Survey Methodology*, 10, 97-118.

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

Hidiroglou, M. and Patak, Z. (2006). An Application to the Canadian Retail Trade Survey, Journal of Official Statistics, 22, 71-80.

Kott, P.S. (2011a). A Nearly Pseudo-optimal Method for Keeping Calibration Weights from Falling Below Unity in the Absence of Nonresponse or Frame Errors, *Pakistan Journal of Statistics*, 27 (4), 391-396.

Kott, P.S. (2011b). WTADJX Is Coming: Calibration Weighting in SUDAAN When Unit Nonrespondents Are Not Missing at Random and Other Applications, *Proceedings of the ASA Survey Research Methods Section* (available online).

Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models, in *Handbook of Statistics_29B: Sample Surveys: Inference and Analysis*, D. Pfeffermann, & C.R. Rao (Eds.), New York: Elsevier.

Kott, P.S. and Liao, D. (2012). Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-weighting Routine, *Survey Research Methods*, 6, 2, 105-11.

Rao, J.N.K. (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimations Stage, *Journal of Official Statistics* 25, 1-21.

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

RTI (2012). *SUDAAN Language Manual, Release 11.0.* Research Triangle Park, NC: RTI International.

US Department of Health and Human Services (2011). *Drug Abuse Warning Network (DAWN), 2008* (Computer file of survey conducted by the Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality). Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Witt, M. (2010). Estimating the R-Indicator, Its Standard Error, and Other Related Statistics with SAS and SUDAAN, Proceedings of the American Statistical Association, Section on Survey Research Methods (available online).