Journal of Official Statistics, Vol. 19, No. 3, 2003, pp. 265–272

A Practical Use for Instrumental-Variable Calibration

Phillip S. Kott[1]

This article describes a simple scenario where defining an instrumental variable is helpful for computing calibration weights (i.e., weights that satisfy the specified calibration equation yet are asymptotically identical to the inverse selection probabilities). The implicit model is simple regression with an intercept. The choice of instrumental variable can reduce the possibility that any calibration weight will be less than one.

Key words: Remainder weights; positive weights; randomization consistent; population mean; remainder mean.

I. Introduction

Recently, Estevao and Särndal (2000) introduced a ''functional form'' calibration estimator for $T = \sum_U y_k$, where U is a population of N elements, with the following form:

$$t_{CALF} = \sum_{k \in S} w_k y_k \tag{1}$$

where S in the sample,

$$w_k = a_k + \left(\sum_{i \in U} \mathbf{x}_i - \sum_{i \in S} a_i \mathbf{x}_i \right) \left(\sum_{i \in S} q_i \mathbf{z}_i' \mathbf{x}_i \right)^{-1} q_k \mathbf{z}_k' \tag{2}$$

$a_k = 1/\pi_k \geq 1$ is the original sampling weight for element k, $\mathbf{x}_k$ is a row vector of J auxiliary variables associated with k, q_k is an arbitrary constant, and $\mathbf{z}_k$ is a row vector of J instrumental variables, some of which may also be components of $\mathbf{x}_k$. This assumes that $\sum_S q_i \mathbf{z}_i' \mathbf{x}_i$ is invertible. The w_k are called ''calibration weights'' because they satisfy the calibration equation, $\sum_U \mathbf{x}_k = \sum_S w_k \mathbf{x}_k$.

It is easy to show that t_{CALF} is an unbiased estimator for T under the model $y_k = \mathbf{x}_k \beta + \varepsilon_k$, where $E(\varepsilon_k | \mathbf{x}_k) = 0$. Moreover, t_{CALF} is randomization consistent under mild conditions, which we assume here to hold. Finally, under those same conditions and some equally mild restrictions on the variance structure of the ε_k, the anticipated variance (model expected randomization mean squared error) of t_{CALF} is asymptotically invariant to the choice of q_k and $\mathbf{z}_k$.

This estimator is an interesting, if not new, generalization of the standard GREG made

[1] National Agricultural Statistics Service, Research Division, 3251 Old Lee Highway, Fairfax, VA 22050, U.S.A.
Acknowledgments: The author thanks an associate editor and two referees for their useful comments on a earlier draft of this article.

popular by Särndal et al. (1992). An earlier version of t_{CALF} can be found in Brewer et al. (1988), although not in calibration form. In practice, it is not obvious why one would contemplate using a vector for $\mathbf{z}_k$ other than $\mathbf{x}_k$ itself, the usual GREG formulation. As for q_k, it is frequently set equal to a_k. Brewer (1999), however, has argued that setting $q_k = a_k - 1$, the *remainder weight*, more often returns a set of calibration weights where $w_k \geq 1$ for all elements in the sample. Many find this a desirable property since then each sample element can be thought of as at least representing itself.

In this brief article, we will consider the scenario where $\mathbf{x}_k = (1, x_k)$, and the x_k vary within the population and the sample (so matrices are invertible when need be). Using the remainder weights for the q_k helps assure that all $w_k \geq 1$ (for those k in the sample). The addition of a well-chosen $\mathbf{z}_k$, introduced in Section 2, makes that property even more likely. Section 3 discusses the more modest goal of finding a complete set of positive calibration weights. Section 4 contains a small empirical investigation. The discussion in Section 5 concludes that a particular choice of the $\mathbf{z}_k$ produces a set of calibration weights with all $w_k \geq 1$ if any such set exists except in an unusual circumstance.

2. The Instrument and Its Calibration Weights

Let $\mathbf{z}_k = (1, z_k)$, where $z_k = 1$ when $x_k \geq A$, and $z_k = -1$ otherwise. A can be anywhere within the range of the x_k. Our rationale in choosing z_k is to lessen the effect of a large $|x_k - A|$ on w_k. Although we use the term ''instrumental variable'' from the econometrics literature (see Johnston 1972, pp. 278–281) to describe z_k, its purpose in this context is very different from that employed by econometricians.

As noted in the introduction, we also let $q_k = a_k - 1$. This limits the size of $|w_k - a_k|$ when a_k is near 1.

Let S_1 be that part of the sample for which $z_k = 1$, and S_2 be the complement of S_1 within the sample. Let m_R be the remainder-weighted mean of the x_k in S, m_{R1} be the remainder-weighted mean of the x_k in S_1, and m_{R2} be the remainder-weighted mean of the x_k in S_2. Let $\hat{R}_1$ be the sum of the remainder weights in S_1, $\hat{R}_2$ be the sum of the remainder weights in S_2, and $\hat{R} = \hat{R}_1 + \hat{R}_2$. Let M_R be the mean value of all x_k in $\mathbb{R} = U - S$, and R be the size of $\mathbb{R}$. Not surprisingly, R is estimated by $\hat{R}$. Under many designs, the two are identical.

Inspecting Equation (2) one can see that the calibration weights are invariant to linear transformations of $\mathbf{z}_k$ or $\mathbf{x}_k$; that is, $\mathbf{z}_k$ or $\mathbf{x}_k$ can be replaced by $\mathbf{z}_k\mathbf{H}$ where $\mathbf{H}$ is any non-singular $J \times J$ matrix without it affecting the result. This means that each component of $\mathbf{z}_k$ can be replaced by a linear combination of other components as long as the combination is the same for all k.

Let us replace x_k in $\mathbf{x}_k$ by $x_k - m_R$. This changes the second component of $\mathbf{x}_k$ to the difference between the original second component and m_R times the first. In $\mathbf{z}_k$, let us replace the 1 by $1/\hat{R}$ times 1, and then replace the second component, z_k, by $(1/\hat{R}_1 - 1/\hat{R}_2)/2$ times itself added to $(1/\hat{R}_1 + 1/\hat{R}_2)/2$ times the original first component (i.e., 1). This effectively replaces z_k with $1/\hat{R}_1$ when $k \in S_1$, and with $-1/\hat{R}_2$ when $k \in S_2$. As a consequence of all these changes, the 2×2 matrix $\sum_S q_i \mathbf{z}_i' \mathbf{x}_i$ becomes diagonal. Its upper left hand corner contains a 1, and its lower right the value $m_{R1} - m_{R2}$. It is important to note that A must be so located within the range of sample

values of x_k that neither S_1 nor S_2 is empty. Otherwise, either R_1 or R_2 would be zero, and the transformation of $\mathbf{z}_k$ above would not be possible.

A little manipulation reveals

$$w_k = a_k + \left(N - \sum_{i \in S} a_i\right)(a_k - 1)/\hat{R} + (m_{R1} - m_{R2})^{-1} \times \left[\sum_{i \in U} (x_i - m_R) - \sum_{i \in S} a_i(x_i - m_R)\right](a_k - 1)c_k \quad (3)$$

where $c_k = 1/\hat{R}_1$ when $k \in S_1$, and $c_k = -1/\hat{R}_2$ otherwise. Observe that for sample designs where $\sum_S a_i = N$, Equation (3) has a much simpler form:

$$w_k = a_k + (m_{R1} - m_{R2})^{-1}\left[\sum_{i \in U} x_i - \sum_{i \in S} a_i x_i\right](a_k - 1)c_k$$

Continuing from Equation (3),

$$\begin{aligned} w_k &= a_k + \left(R - \sum_{i \in S} [a_i - 1]\right)(a_k - 1)/\hat{R} \\ &\quad + (m_{R1} - m_{R2})^{-1}\left[\sum_{i \in \mathbb{R}} (x_i - m_R) - \sum_{i \in S} (a_i - 1)(x_i - m_R)\right](a_k - 1)c_k \\ &= 1 + (a_k - 1) + (R - \hat{R})(a_k - 1)/\hat{R} + (m_{R1} - m_{R2})^{-1}R(M_R - m_R)(a_k - 1)c_k \\ &= 1 + (a_k - 1)[(R/\hat{R}) + (m_{R1} - m_{R2})^{-1}R(M_R - m_R)c_k] \\ &= 1 + (a_k - 1)R(m_{R1} - m_{R2})^{-1}[(m_{R1} - m_{R2})/\hat{R} + (M_R - m_R)c_k] \\ &= 1 + (a_k - 1)(R/\hat{R}_1)(m_{R1} - m_{R2})^{-1}(M_R - m_{R2}) \quad \text{when } k \in S_1 \quad (4.1) \\ &= 1 + (a_k - 1)(R/\hat{R}_2)(m_{R1} - m_{R2})^{-1}(m_{R1} - M_R) \quad \text{when } k \in S_2 \quad (4.2) \end{aligned}$$

This last step uses the equality $\hat{R}_1 m_{R1} + \hat{R}_2 m_{R2} = \hat{R} m_R$.

It is easy to see that w_k in Equations (4.1) and (4.2) will be 1 or larger as long as $m_{R2} \leq M_R \leq m_{R1}$. Now, m_{R1} is a randomization consistent estimator of the mean of the x_k values in $\mathbb{R}$ that are larger than or equal to A, while m_{R2} is a randomization consistent estimator of the mean of the x_k values in $\mathbb{R}$ that are less than A.

In principle, A can be anywhere within the range of the x_k in U. In practice, it makes sense to put it somewhere in the ''center'' of the distribution. Although the population median seems a reasonable choice, the population mean proved more effective in the small empirical example to be discussed in Section 4. In Section 5, we see that setting $A = M_R$ will usually find calibration weights with all $w_k \geq 1$ if such a set exists. Whatever the choice for A, it should to be made before one looks at the sample values, so that the randomization consistency of t_{CALF} is assured, although A can be a predetermined function of the sample values.

3. Sample Weights Versus Remainder Weights

One can think of the conventional ratio estimator as having the same form of t_{CALF} in Equation (1) with $\mathbf{x}_k = x_k$, $\mathbf{z}_k = 1$, and $q_k = a_k$, the original sample weight of element

k. Brewer (1979) proposed a variant of the ratio with $q_k = a_k - 1$, the remainder weight. The name derives because $\sum_S (a_k - 1)y_k$ is an estimator for $\sum_{\mathbb{R}} y_i$.

Each of the calibration weights under the conventional ratio formulation must be positive as long as all $x_i \geq 0$ and one sample element has a positive x-value, since $w_k = \left[\sum_U x_i \Big/ \sum_S a_i x_i\right] a_k$. Brewer's variation assures more. No calibration weight will be less than 1, since $w_k = 1 + \left\{\sum_{\mathbb{R}} x_i \Big/ \sum_S [a_i - 1]x_i\right\}[a_k - 1]$, and $a_k \geq 1$. Note that $[a_i - 1]x_i$ must be positive for at least one sample element for Brewer's w_k to be defined. That is to say, at least one noncertainty sample element must have a positive x-value.

A similar thing happens in our scenario. Defining z_k as in the previous section but letting $q_k = a_k$ in Equation (2), the interested reader can derive these calibration-weight formulae:

$$w_k = a_k(N/\hat{N}_1)(m_1 - m_2)^{-1}(M - m_2) \quad \text{when } k \in S_1 \tag{5.1}$$

$$= a_k(N/\hat{N}_2)(m_1 - m_2)^{-1}(m_1 - M) \quad \text{when } k \in S_2 \tag{5.2}$$

where $\hat{N}_1(\hat{N}_2)$ is the sum of the a_k in $S_1(S_2)$, $m_1(m_2)$ is the sample-weighted mean of the x_k in $S_1(S_2)$, and M is the mean of the x_k in U. Under this formulation, all the calibration weights are positive when $m_2 < M < m_1$. This is no guarantee, however, that each weight is at least 1.

4. A Small Empirical Investigation

In this section, we investigate equally-weighted samples of size 16 drawn from a very large population, U. The population is so large that the differences between Equations (4) and (5) virtually vanish, and positive calibration weights are effectively always larger than 1. As a consequence, we limit our investigation to whether or not the calibration weights are positive.

The x-values are generated by a chi-squared distribution with 1 degree of freedom. The population mean of the x_k in U is assumed to be 1, the mean of the chi-squared distribution. Likewise, its median is assumed to be 0.455.

Although we are rarely interested in samples of size 16 in practice, this study has instructive value. Moreover, it is not that uncommon to use a separate regression estimator where there are as few as 16 sampled elements per stratum.

Table 1 displays the results of four simulations each based on 1,000 independent sample selections. Each of the simulations compares calibration weights computed using Equation (5) with conventional simple regression weights (i.e., those resulting from Equation (2) with $q_k = a_k$ and $\mathbf{z}_k = (1, x_k)$):

$$w'_k = a_k \left[1 + 16(M - m)(x_k - m) \Big/ \sum_{i \in S} (x_i - m)^2\right] \tag{6}$$

where M is the population mean of the $x_k (= 1)$, and m is the sample-weighted mean of the x_k. In the first simulation, A is set equal to the population mean, 1. It is not possible to compute Equation (5) for five of the samples, because all the sample x-values are less than 1. Consequently, S_1 is empty, and m_1 does not exist. In the other 995 samples, the calibration weights are all positive. By contrast, Equation (6) produces at least one nonpositive weight in 6.7% of the samples.

Table 1. Fractions of 1,000 samples with at least one nonpositive weight

A (for Equation (5))	Using Equation (5)	Using Equation (6)[a]
Population mean	0.005[b]	0.067
Population median	0.036	0.065
Sample mean	0.025	0.060
Sample median	0.087	0.055

[a]These values vary because they are based on different samples.
[b]Calibration weights could not be calculated at all in five samples.

Note that it is conceivable for the largest x-value in a sample to be exactly 1, rendering Equation (5) computable and a calibration weight exactly equal to zero. That did not happen in any of the 1,000 samples.

In the next simulation, Equation (5) is calculated using the population median as A. The calibration weights can always be calculated with each of the 1,000 samples, but some weights are nonpositive in 3.6% of them. Although this is better than using the simple regression weights, it is not as good as setting A equal to the population mean in the first simulation.

Table 1 also displays results from simulations using the sample mean and then the sample median as A. Using the sample median produces nonpositive weights in *more* samples than the conventional regression method. Using the sample mean is much better, but not as good as using the population mean.

According to Table 1, setting A equal to the population mean ($A = 1$) is clearly the best thing to do. Increasing the sample size to 25 has little qualitative effect on the results, except that complete sets of positive calibration weights become more common. On the one hand, using Equation (5) with $A = 1$ produces a positive calibration weight for every sample element in all 1,000 samples (not displayed). On the other, Equation (6) with 25 replacing 16 returns at least one nonpositive calibration weight in only 2.3% of the samples. This is a small fraction, but not zero.

5. Discussion

The population mean works well as A in our simulation because Equation (5) will always return nonnegative weights as long as there is at least one sample element with an x-value greater than or equal to the population mean and at least one sample element with an x-value below the population mean.

For those surveys where the population size (N) is not effectively infinite, Equation (4) can be different from Equation (5). The former was constructed to assure that no calibration weight would be less than 1. If we set $A = M_R$ (the mean x-value among population elements not in the sample), then that will always be the case as long as there is at least one sample element with an x-value larger than or equal to M_R and at least one sample element with an x-value below M_R. This is why Equation (4) with $A = M_R$ is usually preferable to Equation (5) with $A = M$. Nevertheless, under certain unusual conditions, it is possible that Equation (5) will return all positive weights, while Equation (4) will not be computable. This can happen when there is no sample element with an x-value larger than or equal to M_R but there is one with an x-value larger than $M < M_R$.

Suppose $x_k \leq x_{smax} < M_R$ for all elements k in the sample, so that Equation (4) with $A = M_R$ is not computable. It is easy to see that no set of calibration weights satisfying $w_k \geq 1$ exists. Suppose one did. Then $\sum_S (w_k - 1) = R$, and $\sum_S (w_k - 1)x_k = \sum_{\mathbb{R}} x_k = RM_R$. But $\sum_S (w_k - 1)x_k \big/ \sum_S (w_k - 1) \leq \sum_S (w_k - 1)x_{smax} \big/ \sum_S (w_k - 1) = x_{smax} < M_R$. This is a contradiction. An analogous argument applies when $x_k \geq x_{smin} > M_R$ for all elements in the sample. There remains the possibility that the minimum value of x_k in the sample is M_R, so that Equation (4) is not computable (because m_{R2} is undefined) but a set of calibration weights all larger than or equal 1 exists with $w_i = 1$ whenever $x_i \neq M_R$.

It is now easy to see that unless the minimum value of x_k in the sample is M_R, Equation (4) with $A = M_R$ fails to produce a set of calibration weights that are all 1 or larger only when no such set exists. A similar series of arguments can be made about Equation (5) with $A = M$ and nonnegative calibration weights.

Let us consider the case where the minimum x_k in the sample is M_R, and Equation (4) fails to find a desired set of calibration weights even though one exists. Suppose we were willing to look at the sample values, see that the minimum value of x_k in the sample was M_R, and change the definition of z_j so that it equaled -1 when $x_j = M_R$. These steps would violate the randomization consistency of t_{CALF} in Equation (1), but it is instructive to note that Equation (4) would again return calibration weights all equal to 1 or larger unless *all* the x_k in the sample equaled M_R, which contradicts the assumption that there is variability among the x_k in the sample.

Readers of an earlier draft of this article have asked whether there is a randomization-mean-squared-error (MSE) penalty from using the nonstandard values for the q_k and $\mathbf{z}_k$ in Equation (2) suggested in the text. The answer to that question is theoretically elusive when the y_k are not constrained by a model and awaits thorough empirical study. Some insight, however, comes from Brewer (1999), who found no such systematic penalty from using a variant of our q_k in a multivariate setting with real data. In particular, he explored using $q_k = (a_k - 1)/g_k$, where g_k was a measure of size somewhat related to his $\mathbf{x}_k$ vector, and found the results very similar to those from using $q_k = a_k$ in terms of empirical MSE.

This leaves open the question of a possible MSE penalty from using an instrumental variable other than x_k. The author takes some solace in noting that the conventional ratio estimator effectively employs an instrumental variable, as pointed out in Section 3, yet the literature reveals no clear penalty from its use.

A simulation of 6,400 samples was run using the same population structure as in Section 4 with $n = 16$. Each y_k was set equal to $x_k^2/3$, so that the linear model, $y_k = \beta_1 + \beta_2 x_k + \varepsilon_k$ with $E(\varepsilon_k | x_k) = 0$, clearly failed. Conveniently, the population mean of the y_k, the goal of the estimators to be described below, is 1. Note that since the population size is assumed to be very large, our goal switches from estimating the population total of the y_k to estimating their population mean.

Computing weights with Equation (6) and each $a_k = 1/n$ produces the conventional simple regression estimator, which is asymptotically equivalent to the randomization-unbiased difference estimator with the lowest randomization MSE; that is, $t_{DIFF} = \sum_S y_k/n + [M - m]b$, where $M = 1$ is the population mean of the x_k, m is the sampled mean of the x_k, and b is chosen to minimize the randomization variance of t_{DIFF}. (See Särndal et al. (1992), Section 6.8.) Despite this fact, the display in Table 2 shows that

Table 2. Properties of different weighting approaches

Estimator	Empirical bias	Empirical root MSE
Sample mean	0.010[a,b]	0.810[b]
Simple regression	−0.182[b]	0.342[b]
Instrumental-variable regression	−0.122	0.363
Modified simple regression	−0.122	0.518[b]

The simple regression estimates were computed with calibration weights determined by Equation (6) with $a_k = 1/n$.
The instrumental-variable regression estimates were computed with calibration weights determined by Equation (5) with $A = 1$ and $a_k = 1/n$. When a set of all-positive calibration weights did not exist, the weights were set to $1/n$ (as in the sample mean).
The modified simple regression estimate was computed with calibration weights determined by Equation (6) with $a_k = 1/n$. When a set of all-positive calibration weights did not exist, the weights were set to $1/n$.
[a]Not significantly different from 0 at the .05 level.
[b]Significantly different from the instrumental-variable-regression value at the 0.05 level.

the simple regression estimator had a more negative empirical bias (−18.2% in percentage terms) than the instrumental-variable regression estimator resulting from computing the calibration weights using Equation (5) with $A = 1$, the population mean of the x_k (−12.2%). Moreover, the simple regression estimator has only a modestly smaller empirical root mean squared error (0.342) than the instrumental-variable regression estimator (0.363). Both methods had much smaller empirical root MSE's than the sample mean (0.810), although the latter is randomization unbiased.

For those few simulations where Equation (5) failed to yield a set of all-positive calibration weights, the calibration weights were set to $1/n$. When the same policy was followed with Equation (6), as is sometimes done in practice, the empirical bias was greatly reduced (−12.2%) but with a noticeable root MSE penalty (0.518). The table refers to the resulting estimator as the ''modified simple regression estimator.''

It is of some interest to note that the model failure resulting from the choice of y_k coupled with the small sample size (16) caused the simple and instrumental-variable regression estimators to be biased. Nevertheless, both had considerably less MSE than the randomization-unbiased sample mean. One does not need the model to hold, for regression estimation to be effective.

Although we need to be cautious about drawing a conclusion from the modest simulations described above, it appears that determining calibration weights with Equation (5) and A equal to the population mean of the x_k is at least competitive with more conventional alternatives.

6. References

Brewer, K. R. W. (1979). A Class of Robust Sample Designs for Large-Scale Surveys. Journal of the American Statistical Association, 74, 911–915.

Brewer, K. R. W. (1999). Cosmetic Calibration with Unequal Probability Sampling. Survey Methodology, 25, 205–212.

Brewer, K. R. W., Hanif, M., and Tam, S. M. (1988). How Nearly Can Model-Based

Prediction and Design-Based Estimation Be Reconciled? Journal of the American Statistical Association, 83, 128–132.

Estevao, V. M. and Särndal, C.-E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379–400.

Johnston, J. (1972). Econometric Methods, 2nd Edition. New York: McGraw-Hill.

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

Received May 2002
Revised December 2002