Calibration Weighting And Non-Sampling Errors

Chris Skinner

Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, United Kingdom

Summary

Calibration weighting provides an important class of techniques for the efficient combination of data sources. These techniques have been developed under classical sampling assumptions in the absence of non-sampling error. In practice, however, calibration is often used to correct for non-response bias. This paper explores the properties of calibration estimation in the presence of both nonresponse and measurement errors. The ideas are illustrated with a simple example concerning the estimation of the number of sight tests carried out in Great Britain.

1 Introduction

It is often desirable to make use of several data sources when producing statistical estimates. First, a more accurate estimate may be achievable from a combination of sources than from any single source. Second, the presence of common variables in different data sources may lead to incoherence if estimates from the different sources are produced separately.

Calibration estimation (Deville and Särndal, 1992) provides a valuable class of techniques for combining data sources. The basic idea is to use estimates from one set of sources, which may be treated as sufficiently accurate to act as 'benchmarks'. Estimates based on data from a further sample source are then adjusted so as to agree with these benchmarks. The process of adjustment is called 'calibration'. The constraints that the estimates of the benchmarks based on this source should agree with the benchmarks are

called ‘calibration constraints’.
A practical attraction of calibration estimation is that it may be achieved computationally in two straightforward steps. First, the benchmark information is used to construct a simple set of weights for all units in the sample source and then these common weights may be used to adjust any estimates based on this source. Such computation may be implemented in software such as CALMAR, CLAN and GES developed at INSEE, Statistics Sweden and Statistics Canada, respectively.
Simple examples of calibration estimation are provided by ratio estimation and postratification. In the classical case it is assumed that population values are available for an auxiliary variable and that these data are combined with sample data on some survey variable to estimate the mean or total of this variable. In ratio estimation it is assumed that the population total or mean of a continuous auxiliary variable is known. In postratification it is assumed that the population proportions falling into the categories of a discrete auxiliary variable are known. Classical sampling textbooks, such as Cochran (1977), demonstrate how such techniques may lead to improved precision, i.e. reduction of the variance of the sampling error, compared to estimators which make no use of the auxiliary information.
The classical framework assumes that the auxiliary population information is correct and that the sample source is obtained using a known probability sampling scheme. In particular it is assumed that there are no sources of non-sampling error such as non-response, noncoverage or measurement error. Within this framework, Deville and Särndal (1992) demonstrate that a wide class of calibration estimators are consistent and that the large-sample efficiency of many calibration estimators is equal to that of a generalised regression estimator. The choice between alternative calibration estimators may then depend more on non-asymptotic considerations such as the desirability for weights to be non-negative and not too variable (e.g. Chambers, 1996). Alternative calibration estimators might also be considered to improve efficiency (Montanari, 1987, 1998; Rao, 1994) or to reduce mean squared error by relaxing the calibration constraints (Chambers, 1996).
In this paper, we go beyond the classical framework and allow, more realistically, for non-sampling error. Calibration estimation is, indeed, widely used to reduce bias from non-response (e.g. Bethlehem, 1988; Fuller et al, 1994) and noncoverage and some have argued (e.g. Kalton and Flores-Cervantes, 1998) that, in practice, this is usually the main purpose of such adjusted estimation. In this paper we shall allow for both nonresponse and measurement error and consider to what extend calibration estimation leads to improved estimation quality.

2 An Example: Estimating the Number of Sight Tests

To provide a simple illustration of the ideas developed in this paper we consider a simple example. The British Department of Health wishes to produce estimates of the number of sight tests carried out in Great Britain each year. Sight tests may be conducted by either Optometrists or Ophthalmic Medical Practitioners, who will be referred to jointly as ‘opticians’ for simplicity. A principal data source for estimates of numbers of sight tests is the Sight Tests Volume and Workload Survey which involves the completion of questionnaires and diary sheets by a sample of opticians. Sight tests may be divided between private tests and tests which are funded by the government’s National Health Service (NHS), part of the Department of Health. Roughly equal numbers of each type of test have been conducted recently (in 1994/95 it was estimated that 50.3% of tests were private). Administrative records are also kept of the number of NHS tests paid for

and these provide an alternative data source for producing estimates of the total number of NHS tests carried out. This information may also be used for calibration estimation of the numbers of all tests. For example, in one quarter of 1993-94 the number of NHS tests estimated from administrative sources was

T_x = 1.672 M (M = million)

The corresponding survey estimate was

$\hat{T}_{xs}$ = 1.579 M

This estimate may be expressed as $\hat{T}_{xs} = \sum d_i x_i$ where d_i is the basic survey weight, x_i is the number of NHS tests conducted by a sampled optician and the sum is over the sample. A ratio estimate of the total of a variable y_i then takes the form

$$\hat{T}_y = \sum w_i y_i , \quad \text{where } w_i = d_i T_x / \hat{T}_{xs} = 1.059 d_i$$

and the sum is again over the sample. Note that the ratio estimate of the number of NHS tests is

$$\hat{T}_x = \sum w_i x_i = \left(\sum d_i x_i\right) T_x / \hat{T}_{xs} = \hat{T}_{xs} T_x / \hat{T}_{xs} = 1.672M.$$

Thus, the ratio estimation procedure constrains survey estimates so that the estimate of the total number of NHS tests is equal to the benchmark value obtained from administrative sources. This defines the calibration constraint.

The basic survey estimate of the number of private sight tests was 1.578M and so the corresponding ratio estimate is 1.059 × 1.578 = 1.670M. Because the weighting is linear, the ratio estimate of the number of all sight tests is 1.672 + 1.670 = 3.342M.

3 Calibration Methods

Suppose that the aim is to estimate the total T_y of a (scalar) variable y_i across units i in a population P. Data are assumed available from two sources. First, values y_i^s and x_i^s are recorded in a sample survey for units in a set r of respondents. If y_i is measured without error in the survey then we simply write

$y_i^s = y_i$. Otherwise y_i^s denotes the measured value, y_i the true value and $y_i^s - y_i$ the measurement error. In general x_i^s is a 1 × J vector of values.

A second source of data provides the vector T_x of population totals of the 1 × J vector x_i^a

It is useful to distinguish two cases:

Case 1: $x_i^s = x_i^a$ for all responding i, that is both data sources lead to identical measurements;

Case 2: $x_i^s \neq x_i^a$ for some responding i, that is there is some variation between the measurements obtained in the sources.

These two cases will be considered further in Section 5. In neither case will it be necessary to introduce notation for true values corresponding to x_i^s and x_i^a. In the classical framework it is assumed that $y_i^s = y_i$, $x_i^s = x_i^a = x_i$, say, for all units i in the respondent set r, which is drawn from the population using a probability sampling scheme with known inclusion probabilities π_i. In this case a general calibration estimator of T_y is defined by

$$\hat{T}_y = \sum_r w_i y_i$$

where the weights w_i are chosen to minimise $\sum_r G_i(w_i, d_i)$, which measures the distance between the w_i and the design weights $d_i = \pi_i^{-1}$, subject to the following J

calibration constraints being obeyed:

$$\sum_r w_i x_i = T_x \ .$$

Different choices of the functions G_i will lead to different estimators. The choice $G_i(w_i, d_i) = (w_i - d_i)^2 / 2d_i q_i$ leads to the generalised regression (GREG) estimator with a closed form expression for the optimal w_i, namely

$$w_i = d_i \left[1 + q_i x_i \left(\sum_r d_i q_i x_i^T x_i \right)^{-1} \left(T_x - \hat{T}_{xs} \right)^T \right],$$

where $\hat{T}_{xs} = \sum_r d_i x_i$ and the q_i are constraints to be specified, usually dependent upon an assumed variance function for y_i given x_i.

An important result of Deville and Särndal (1992) is that, within the classical framework and under further regularity conditions, estimators $\hat{T}_y$ with different choices of function G_i are asymptotically equivalent and, in particular, are all consistent for T_y with the same asymptotic variance. Thus, the choice of function G_i will depend more on considerations about the ease of computation of the w_i or about the 'small sample' properties of the w_i. One of these properties concerns the occurrence of negative w_i. In the classical asymptotic framework, the weights w_i will converge to the design weights as the sample size increases and (under weak conditions) the probability that any w_i is negative will tend to zero (since all d_i are positive). In finite samples, however, some w_i may be negative and G_i might be chosen in order to avoid this possibility.

Another 'small sample' property is that the variance of the calibration estimator can increase when there are very many calibration constraints and so some selection of constraints (Nascimento Silva and Skinner, 1997) or weakening of these constraints to the requirement that they only hold approximately (Chambers, 1996) may be desirable.

4 Nonresponse

In practice most surveys are subject to nonresponse and many sampling frames from which the sample is drawn are subject to noncoverage. In this case, if auxiliary population information is available then calibration estimation may be used to attempt to reduce bias. Some adaption of the estimator will be necessary, however, since the inclusion probabilities will generally be unknown and some may be equal to zero.

As in the classical framework we assume $y_i^s = y_i$ and $x_i^s = x_i^a = x_i$, say. The basic survey weights d_i may now involve some adjustment of the reciprocals of the sampling inclusion probabilities π_i. One possibility, for example, would be to set $d_i = (n/r)\pi_i^{-1}$, where n is the intended sample size and r the number of respondents. In fact the choice of adjustment may not have a great effect on bias (see Bethlehem, 1988; Lundström, 1997). For example, the bias of the GREG estimator may be shown (following the approach of Lundström, 1997) to be approximately

$$\sum_P (y_i - x_i B), \quad \text{where} \quad B = \left(\sum_P d_i \theta_i q_i x_i^T x_i \right)^{-1} \sum_P d_i \theta_i q_i x_i^T y_i$$

and θ_i is the probability of unit i being included in the data, that is the sampling inclusion probability π_i multiplied by the probability that a unit responds given that it is sampled. This result assumes the q_i may be

expressed as a linear combination of the vector x_i, i.e. $q_i = cx_i^T$. Now if the nonresponse and sampling are ignorable given x and if y has a linear regression on x, $E_\xi(y_i \mid x_i) = x_i\beta$, then the model expectation of B is equal to β and is free of the choice of adjustment used to define d_i. In any case it is clear that B is unaffected by the multiplication of d_i by a constant such as (n/r).

Example: A simplified version of the sight test example from Section 2 follows. The sampling design involves (mildly) disproportionate stratification and the basic survey-based estimator is

$$\hat{T}_{ys} = \sum_r d_i y_i ,$$

where d_i is the ratio of the population size to the number of respondents in the stratum containing i. The ratio estimator is

$$\hat{T}_y = \left(\sum_r d_i y_i \Big/ \sum_r d_i x_i \right) \sum_P x_i$$

Under nonresponse the approximate expectations with respect to the response and sampling mechanisms are

$$E(\hat{T}_{ys}) = \sum_P \theta_i d_i y_i \quad , \quad E(\hat{T}_y) = \left(\sum_P \theta_i d_i y_i \Big/ \sum_P \theta_i d_i x_i \right) \sum_P x_i .$$

Clearly both estimators are unbiased if $d_i = \theta_i^{-1}$, which would occur here if nonresponse is completely at random within strata. A common linear regression through the origin across strata for y = number of all sight tests (or number of private tests) on x = number of NHS tests $\left(E(y_i \mid x_i) = x_i\beta\right)$ provides a good fit to the respondent data. Assuming this model applies to both nonrespondents and respondents, so that nonresponse is ignorable given x, the biases of the estimators with respect to both the model and the response and sampling mechanisms are approximately

$$E(\hat{T}_{ys} - T_y) = \sum_P (\theta_i d_i - 1) x_i \beta \quad , \quad E(\hat{T}_y - T_y) = 0$$

For this reason, the ratio estimator may lead to reduced bias if the rate of nonresponse varies within strata according to the size variable x_i, but not otherwise with respect to y_i.

A further problem with nonresponse is concealed 'overcoverage'. The sampling frame consists of a set of registers of opticians who are qualified to undertake sight tests. Some of the opticians on these registers are not in fact practising, for example because they have retired or they are devoting their work to other activities, so that their values of y_i and x_i will be zero. A greater proportion of nonrespondents than respondents may be expected to fall into this group. This will tend to lead to upward bias in the survey-based estimator $\hat{T}_{ys}$. Because of this, some attempt is made to use estimated practising rates to adjust the population sizes used to calculate the d_i. Problems in estimating practising rates may, however, lead to further biases. Following a similar argument to that above, the ratio estimator may be expected to be subject to less bias, even without adjusting for estimated practising rates, provided it is reasonable to suppose that there is little correlation between any variable regression slopes β in different strata and between any different practising rates in different strata.

The calibration constraints imply that the calibration estimator will have zero error for any choice of G_i function if y_i is a linear combination of the elements of x_i. In general, however, the asymptotic bias of the calibration estimator will depend on the G_i function when nonresponse is present, unlike the classical case. The classical asymptotic

distribution fails to hold, in particular because $T_x - \hat{T}_{xs}$ no longer converges to a zero vector in general. If the survey weights are taken to be proportional to the reciprocals of the sample inclusion probabilities then one condition under which the calibration estimator is approximately unbiased is when the probability of response ϕ_i is equal to d_i/w_i. Clearly this condition varies for different choices of G_i function. For example, Kalton and Maligalig (1991) note that this holds if the ϕ_i take a multiplicative form for a raking ratio estimator.

A further general consequence of nonresponse is that the weights w_i will not converge to the original weights d_i as the sample size increases. For example, for the GREG estimator

$$w_i \rightarrow \tilde{w}_i = d_i(1 + q_i x_i \psi),$$

where $$\psi = \lim\left[\left(\sum_r d_i q_i x_i^T x_i\right)^{-1}\left(T_x - \hat{T}_{xs}\right)^T\right]$$

In particular, the presence of nonresponse may be expected to lead to negative weights much more frequently.

Yet another consequence of nonresponse is that the variance of the calibration estimator will be dependent on the G_i function and revised methods of variance estimation need to be considered (Lundström, 1997).

5 Measurement Error

The possibility of measurement error in either the survey variables or the benchmark variables is now considered. We first recall the distinction made in Section 3.

Case 1 ($x_i^s = x_i^a$ for all responding i).

This arises for example in business surveys where the same source, a business register, is used to derive both the values x_i^s for the respondents and the x_i values upon which the auxiliary vector X is based.

Case 2 ($x_i^s \neq x_i^a$ for all responding i).

This arises when x_i^s is obtained as a survey measurement, e.g. the number of NHS sight tests reported by an optician, and the auxiliary total x_i^a is based on a different data source, e.g. administrative records.

These cases are treated in turn.

Case 1 ($x_i^s = x_i^a$)

In this case measurement error in the auxiliary variables will not tend to introduce bias into the calibration estimator but only lead to a loss of precision. For example, suppose that x_i^a is the number of employees in a firm as recorded on a business register. If this figure becomes out of date and thus subject to error, it may become less correlated with the y_i variables of interest and thus reduce precision in calibration estimates. The option of replacing the register employment by a figure obtained in the survey may reduce the variance of the estimator but is likely to introduce bias (see Case 2)).

The presence of zero mean measurement error in a continuous y_i variable may be expected to lead to a similar loss of precision in either the survey estimator or the calibration estimator. Similarly the presence of measurement error with non-zero mean may be expected to lead to a similar bias in either estimator.

Case 2 ($x_i^s \neq x_i^a$)

In this case, the use of calibration estimation may introduce bias and be undesirable. Steel (1997) provides an example on the use of Labour Force Survey data in Great

Britain to estimate the number of unemployed, according to the ILO definition. A possible auxiliary variable for which the population total is known is the binary variable x_i, indicating whether or not a person claims unemployment benefit. Postratification by x_i may improve precision, reducing the variance by 35%. The problem is that claimant status is underreported in the LFS so that $x_i^s \neq x_i^a$. As a result postratification may lead to the estimate of ILO unemployment being 12% too high. This bias effect far outweighs the gain in precision in term of mean squared error.

Calibration estimation should not, however, be dismissed immediately in Case 2. It is possible for it to reduce bias if y_i is also subject to error. Consider, for example, the sight test example, where y_i^s = number of private sight tests recorded in survey, x_i^s = number of NHS sight tests recorded in survey,

x_i^a = number of NHS tests recorded in administrative sources. Suppose that both y_i^s and x_i^s are subject to the same kind of underreporting, common in diaries, so that $y_i^s = 0.9\ y_i$ and $x_i^s = 0.9\ x_i^a$ (the administrative source here is assumed error free). Then the usual survey estimator will be biased downwards by 10%. But, the ratio estimator will be approximately unbiased because the underreporting in y_i^s and x_i^s cancel each other out.

Of course, the observed difference reported in Section 2 of $\hat{T}_{xs} = 1.579$M and T_x = 1.672M could be due to error in the auxiliary figure T_x rather than underreporting in the survey and so ratio estimation need not necessarily reduce bias here.

Finally, it is worth noting that in some circumstances, Case 2 can be changed to Case 1 by matching records between data sources and reconciling the x_i^s and x_i^a values. Often this will be difficult or impossible, however, for practical or confidentiality reasons.

6 Testing for Non-Sampling Bias

Let d_i be the basic survey weight, possibly adjusted by some nominal response probability, and let $\hat{T}_{xs} = \sum d_i x_i$ be the corresponding basic survey estimator of the vector of benchmarks T_x. If $\hat{T}_{xs}$ is asymptotically unbiased for T_x then calibration will not affect asymptotic bias and, in particular, not introduce it if none is present. It may therefore be useful to check whether this condition holds. A natural approach is to use a Wald test based on the statistic

$$X_w^2 = \left(\hat{T}_{xs} - T_x\right) V^{-1} \left(\hat{T}_{xs} - T_x\right)^T ,$$

where V is an estimate of the covariance matrix of $\hat{T}_{xs}$. If X_w^2 lies in the critical region of the chi-squared distribution with J degrees of freedom then the hypothesis that $\hat{T}_{xs}$ is asymptotically unbiased for T_x may be rejected. If this is the case then a variety of possible explanation are possible. First, nonresponse or noncoverage may be differential with respect to x_i, in a way which is not controlled for by any initial adjustment of the d_i weights. Second, there may be a systematic difference between the survey measures x_i^s and the auxiliary measures x_i^a (Case 2 of Section 5), that is measurement bias may exist.

In order to distinguish between these rival explanations it is likely to be necessary to make judgements which go beyond the data. In the sight tests example, both possibilities seem plausible as initial hypotheses. Thus, it is plausible that nonresponse is differential by workload and it is plausible that underreporting may occur in diary data. Follow-up of the survey process seems necessary to investigate these hypotheses further.

7 Conclusions

Non-sampling errors are critical determinants of the quality of official statistics. They may interact with calibration adjustments in complex ways and many of the classical properties of calibration estimation no longer apply. We have outlined circumstances in which calibration estimation may be expected to reduce bias from both nonresponse, noncoverage and measurement error. We have also noted one instance where calibration may introduce severe bias, i.e. when the measurement of the auxiliary variables differs systematically between the survey and benchmark data sources. More research is needed to investigate the properties of calibration estimates in the presence of non-sampling errors, to investigate possible adjustments to calibration estimators e.g. using multiphase information, to consider variance estimation of calibration estimators in the presence of nonsampling errors and to consider strategies for the choice of calibration constraints and estimators in the presence of non-sampling errors. See Lundström (1997) and Skinner and Nascimento Silva (1997) for some further ideas on these issues.

References

Bethlehem, J. G. (1988) Reduction of nonresponse bias through regression estimation. J. Off. Statist. 4, 251 – 260

Chambers, R. L. (1996) Robust case-weighting for multipurpose establishment surveys. J. Off. Statist. 12, 3 – 32

Cochran, W. G. (1997) Sampling Techniques (3rd Ed.) New York: Wiley.

Deville, J. C. and Särndal, C. E. (1992) Calibration estimation in survey sampling. J. Amer. Stat. Assoc. 87, 376 – 382.

Fuller, W. A., Loughin , M. M. and Baker H. D. (1994) Regression weighting in the presence of nonresponse with application to the 1987 – 1988 nationwide Food Consumption Survey. Survey Methodology, 20, 75 – 85

Kalton, G. and Flores-Cervantes, I. (1998) Weighting Methods. In A. Westlake, J. Martin, M. Rigg,
C. Skinner Eds. New Methods for Survey Research Association for Survey Computing,Bucks, U.K.

Kalton, G. and Maligalig, D. S. (1991) A comparison of methods of weighting adjustment for nonresponse, Proceedings of the 1991 Annual Research Conference US Bureau of the Census, Arlington VA, 409 – 428

Lundström, S. (1997) Calibration as a standard method for the treatment of noonresponse. Doctoral dissertation, Department of Statistics, Stockholm University.

Montanari, G. E. (1987) Post-sampling efficient QR – prediction in large scale surveys. Int. Statist. Rev., 55, 191 – 202

Montanari, G. E. (1998) On regression estimation of finite population means. Survey Methodology, 24, 69 – 77.

Nascimento Silva, P. L. d. and Skinner, C. J. (1997) Variable selection for regression estimation in finite populations. Survey Methodology, 23, 23 – 32.

Rao, J. N. K. (1994) Estimating totals and distribution functions using auxiliary information at the estimation stage. J. Off. Statist., 10, 153 – 165

Skinner, C. J. and Nascimento Silva, P. L. d. (1997) Variable selection for regression estimation in the presence of nonresponse. Proc. Surv. Res. Sect. Amer. Stat. Assoc.

Steel, D. (1997) Producing monthly estimates of unemployment and employment according to the International Labour Office definition (with discussion). J. Roy. Statist. Soc., Series A, 160, 5 – 46.