Intelligent Calibration

R. L. Chambers and C. J. Skinner
Department of Social Statistics, University of Southampton
Highfield, Southampton, UK SO17 1BJ
rc@alcd.soton.ac.uk and cjs@alcd.soton.ac.uk
Suojin Wang
Department of Statistics, Texas A&M University
College Station, Texas, USA 77843
sjwang@stat.tamu.edu

1. Introduction

A standard requirement of many sample weighting systems is that the weights be calibrated against known population totals of benchmark variables. That is, they should be capable of exactly reproducing these totals when applied to the survey values of the corresponding variables. The first explicit assessment of calibration as an estimation methodology was carried out by Deville and Särndal (1992), although the idea of calibration is implicit in many of the model-based/model-assisted sample weighting methods that have been developed since the 1970's. The results obtained by these authors showed that a wide class of calibration estimators are design consistent and asymptotically equivalent to the generalised regression estimator. Chambers (1997) developed a model-based approach to the problem based on a linear model framework that allows comparison of a number of calibrated (and quasi-calibrated) sample weighting methods that have been suggested in the literature. In turn, this framework suggests a number of ways sample weighting methods can be improved and extended, in the sense of defining more accurate and more robust estimators of population quantities.

Calibration of sample weights with respect to a specified set of benchmark variables can lead to estimates that have desirable statistical properties; e.g. unbiasedness under a linear population model defined in terms of these variables, increased efficiency if these variables define the "correct" model for the survey variable and robustness to model misspecification if the calibrated weights are also "close" to those associated with a nonparametric estimator of the population quantity of interest. However, calibration comes at a price - increased variance. A variety of methods for minimising this "price" are available; for example Bardsley and Chambers (1984) describe a "quasi-calibration" methodology involving the use of ridge methods to trade off what can be a substantial variance increase under calibration for a small "slippage" in meeting the calibration constraints. In contrast, Silva and Skinner (1997) develop a variable selection strategy for regression estimation and empirically demonstrate that calibration to a small number of "appropriate" benchmark variables can lead to gains in efficiency compared with calibration against a much larger set of benchmarks, many of which are unrelated to the survey variable of interest. Thus a survey weighting system which requires the same sample weight be used for every survey variable, and consequently uses weights that are calibrated with respect to all possible benchmarks associated with these variables, will be less efficient than "variable specific" weights that are calibrated on only those benchmarks that are related to the particular survey variable. Commonly observed consequences of this overcalibration are negative or extreme weights.

Comparatively little work has been carried out on how to balance these conflicting requirements. In a sense, what is required is an "intelligent" approach to calibration, which trades off the possible "biases" that can occur if not all benchmarks are used in calibration against the possible "variances" that arise from overzealous calibration.

2. Calibration for Non-Response
A common use of calibration is to adjust for differential nonresponse. A simple example of this is ratio estimation where an initial weighted estimate of the population total based on respondent data is adjusted for the difference in the nonrespondent and respondent distributions of an auxiliary variable by scaling the sample weights of the respondents so that the corresponding weighted total for the auxiliary variable based on the respondents alone is identical to the known population total of this variable.

The main justification for calibration in the case of nonresponse is the assumption that this nonresponse is ignorable conditional on the values of the auxiliary variables. If this is true, two population units with the same values of these auxiliary variables can be assumed to have values for the study variable that are drawn from the same distribution. If we know the value of the study variable for one unit but not the other, then we can use this known value as a predictor of the other unknown value. Provided the sampling method is also similarly ignorable, this result holds irrespective of whether the unit whose value is unknown is a nonresponding sample unit or a nonsampled unit. Consequently, all we need to do is to model the relationship between the study variable and the auxiliary variables in the respondent data. This “respondent” model can be used to predict the values of the study variable where they are "missing" - no distinction is made between missingness through nonresponse and missingness through exclusion from the sample. If furthermore this model is linear in the auxiliary variables, then standard prediction theory tells us that an unbiased predictor of the total of these missing values of the study variable is given by a weighted sum of the respondents’ values for this variable, where the weights are calibrated to the auxiliary variables in the model. Consequently choice of calibration constraints is synonymous with an implicit linear model specification for the regression of the survey variable on the auxiliary variables defining these constraints. The decision about just what calibration constraints one should impose therefore depends on just what is a reasonable "regression model" for the response variable in terms of the auxiliary variables that are available.

3. What to Calibrate On?

The issue of model choice (and hence calibration constraints) for regression estimation under nonresponse has been investigated by Skinner and Silva (1997). They recommend a stepwise variable selection procedure that identifies the model under which this type of estimator has smallest mean squared error. This is equivalent to identifying the largest subset of the calibration constraints that can be ignored without leading to a significant increase in the mean squared error of the regression estimator. Here we explore the same problem but investigate an alternative approach based on a modified version of the regression estimator that actually achieves minimum mean squared error in this case. In order to finesse complications arising from choice of sampling scheme, we will assume that the only missingness in the data for the survey variable is due to nonresponse. The aim therefore is to estimate (or equivalently, predict) the total of the sampled units’ values for this variable.

Let \( Y \) denote the survey variable of interest, with \( n \) population units sampled. Of these, we only observe \( Y \) for the first \( m \) (the respondent values). However, we do have access to the values of two vector-valued auxiliary variables, say \( X_1 \) and \( X_2 \), for all \( n \) sampled units. Let \( p_k = \dim(X_k) \), \( k = 1, 2 \). Our aim is to predict the total of \( Y \) for the \( n-m \) nonresponding units on the basis of these data. We denote this by \( T \). One way of doing this is to assume a linear regression model for \( Y \) in terms of both \( X_1 \) and \( X_2 \), that is, \( E(Y \mid X_1, X_2) = X_1' \beta_1 + X_2' \beta_2 \), and to predict each nonresponding value of \( Y \) by the best linear unbiased estimate of this conditional expectation. The best linear unbiased estimates of \( \beta_1 \) and \( \beta_2 \) needed for this prediction can be calculated using the respondent data. Since each prediction is a linear function of the respondent \( Y \) values, our estimate of \( T \) will be a weighted sum of these values. We refer to the above model as “A” below, denoting its associated estimate by
\( \hat{F}_A \), and its associated weights by \( w_{Ai} \). It is easy to see that the unbiasedness under A ensures that the \( w_{Ai} \) are calibrated so they recover the sample totals of the components of both \( X_1 \) and \( X_2 \).

Now suppose we actually investigate the relationship between \( Y \) and \( X_1, X_2 \) in the respondent data and discover that the linear regression model is valid, but the coefficient \( \beta_2 \) of \( X_2 \) is not significantly different from zero at some significance level close to zero (e.g., .05). This suggests that it may be preferable to predict the nonrespondent \( Y \)-values on the basis of the reduced model \( E(Y \mid X_1, X_2) = X_1 \beta_1 \), which we refer to as model “B”. This leads to an alternative linear estimate \( \hat{T}_B \), with associated weights \( w_{Bi} \). These weights are calibrated on \( X_1 \) but not, in general, on \( X_2 \). Since \( B \) is nested in \( A \), we follow Skinner and Silva (1997), deciding between these two options on the basis of a comparison of the mean squared errors of \( \hat{T}_A \) and \( \hat{T}_B \) under \( A \). Assuming distinct units are uncorrelated, \( \Delta = \text{MSE}(\hat{T}_B) - \text{MSE}(\hat{T}_A) = (n-m)^2 \tilde{u}_{n-m}^{\dagger} \hat{\beta}_2 \tilde{\beta}_2' - \sigma^2 \tilde{S}_{u,m}^{\dagger} \tilde{u}_{n-m} \) where \( \tilde{u}_{n-m} \) is the vector of column means defined by the matrix \( X_2n-m - X_{1n-m}(X_{1m}X_{1m})^{-1}X_{1m}X_{2m} \) of residuals for the \( n-m \) nonrespondents, and \( \tilde{S}_{u,m} \) is the matrix of crossproducts of these residuals for the \( m \) respondents. The parameter \( \sigma^2 \) denotes the (assumed constant) residual variance of \( Y \) under \( A \). Let \( \hat{\beta}_2 \) and \( \hat{\sigma}^2 \) denote unbiased estimators of \( \beta_2 \) and \( \sigma^2 \) under \( A \). An unbiased estimator of \( \Delta \) is then \( \hat{\Delta} = (n-m)^2 \tilde{u}_{n-m}^{\dagger} \hat{\beta}_2 \hat{\beta}_2' - 2 \hat{\sigma}^2 \tilde{S}_{u,m}^{\dagger} \tilde{u}_{n-m} \).

One option for choosing between \( A \) and \( B \) would be to choose \( A \) when \( \hat{\Delta} > 0 \), and \( B \) otherwise. Note that this is not the same as choosing \( B \) when we cannot reject the hypothesis that \( \beta_2 = 0 \). To see this, observe that in the case \( p_2 = 1 \), \( \hat{\Delta} = (n-m)^2 \tilde{u}_{n-m}^{\dagger} (F - 2) / \hat{\sigma}^2 \), where \( F = \tilde{S}_{u,m}^{\dagger} \hat{\beta}_2^2 / \hat{\sigma}^2 \) is the usual F statistic for testing \( \beta_2 = 0 \). If this test is carried at a 5% significance level, then for large \( m \) we reject this hypothesis (i.e. choose \( A \)) if \( F > 4 \) approximately, that is, when \( \hat{\Delta} > 2(n-m)^2 \tilde{u}_{n-m}^{\dagger} / \hat{\sigma}^2 \). Otherwise we choose \( B \). Thus, the decision to use \( A \) instead of \( B \) on the basis of a test of significance requires “more” evidence against \( B \) than a simple comparison of the value of \( \hat{\Delta} \) would suggest.

Both the simple "zero cut-off" decision rule or the more stringent significance test decision rule for choosing between \( A \) and \( B \) on the basis of \( \hat{\Delta} \) have implications for how the variables defining \( B \) (i.e. \( X_1 \)) are defined. Both rules can be stated as deciding that \( X_2 \) may be ignored (i.e. deciding on \( B \)) when \( \hat{\Delta} < c \), where \( c \) is an appropriately chosen constant. The zero rule has \( c = 0 \), while, when \( p_2 = 1 \), the 5% significance test rule has \( c = 2(n-m)^2 \tilde{u}_{n-m}^{\dagger} / \hat{\sigma}^2 \). In either case, an intuitively appealing approach is to choose \( X_1 \) in order to minimise \( \hat{\Delta} \), for example by conducting a stepwise search procedure that first identifies the variable \( X_2 \) in \( X \) which minimises this statistic, then replaces \( X \) by \( X_1 \) and repeats the procedure, stopping when no \( X_2 \) can be found such that \( \hat{\Delta} < c \). Clearly, at the end of this process one is in the situation described at the start of this section, where \( X \) can be divided into two groups of variables, one group (\( X_1 \)) which is seen as "necessary" for explaining \( Y \) and a second group (\( X_2 \)) which, on the basis of the preceding analysis, seems "unnecessary". The estimator of choice is then \( \hat{T}_B \).

4. Improving On Calibration

Should one stop having decided on an "optimal" \( X_1 \)? We argue the answer to this is no. The fact that the model \( B \) defined by this choice leads to an estimated mean squared error that is smaller than that of model \( A \) does not mean the actual mean squared error satisfies this inequality. However, it is also clear that the variables defining \( X_1 \) and \( X_2 \) in this formulation are not on an equal footing. The former group are considerably more important for explaining \( Y \) than the latter. Using \( \hat{T}_B \) gives the variables in \( X_2 \) no weight in the analysis, while using \( \hat{T}_A \) essentially gives these
variables too much weight. One compromise is to compute $\hat{T}_w = \theta \hat{T}_A + (1 - \theta) \hat{T}_B$, where $\theta$ is chosen to minimise the mean squared error of $\hat{T}_w$. Some algebra leads to this minimising value being $\theta_{opt} = \frac{1}{u_n - m} \frac{\beta_2 \beta_2' \sigma^2 S_{u,m} \beta_2}{(u_n - m)^2 \beta_2' \beta_2 + \sigma^2 S_{u,m} \beta_2}$. This optimal value can be estimated by "plugging in" estimates of $\beta_2$ and $\sigma^2$. In any case, treating $\theta$ as a fixed value, it is clear that $\hat{T}_w$ is a linear estimator, with weights that are calibrated on $X_1$ (since both $\hat{T}_A$ and $\hat{T}_B$ are calibrated on this variable), but not on $X_2$.

Alternatively, one can adopt the “ridge regression” approach to weighting described in Bardsley and Chambers (1984) to define an efficient compromise between $\hat{T}_A$ and $\hat{T}_B$. Here the weights are chosen in order to minimise $(w - 1_m)'(w - 1_m) + \lambda^{-1}(w'X_m - 1_nX_n)'C(w'X_m - 1_nX_n)$, where $1_m, 1_n$ are vectors on ones, $\lambda > 0$ is a suitably chosen scalar (“ridge”) parameter and $C$ is a non-negative definite matrix. The minimising weights are $w_\lambda = 1_m + X_m(\lambda C^{-1} + X_m'X_m)^{-1}X_m'1_{n-m}$. Let $D_{jk}, j,k = 1, 2$ denote the partition of $C^{-1}$ according to the variables making up $X_1$ and $X_2$. By setting $D_{11} = D_{12} = D_{21} = 0$ the ridge weights $w_\lambda$ can be shown to interpolate between $w_{A1}$ ($\lambda = 0$) and $w_{B1}$ ($\lambda = \infty$). Furthermore, setting $D_{22} = S_{u,m}$ forces this interpolation to be linear and leads to $w_\lambda = w_A - (\lambda / (1 + \lambda))(w_B - w_A)$. Consequently, the same argument as above can be used to show that the optimal value of $\lambda$ is $\lambda_{opt} = (1 - \theta_{opt})/\theta_{opt}$, in which case the ridge estimator weights are just $w_\lambda = w_A - (1 - \theta_{opt})(w_A - w_B) = \theta_{opt}w_A + (1 - \theta_{opt})w_B$, and we recover the optimal value of the estimator $\hat{T}_w$. Note that this choice of $\lambda$ differs from the “ridge trace” approach suggested by Bardsley and Chambers (1984), where $\lambda$ is chosen as the smallest value such that all components of $w_\lambda$ are greater than or equal to one.

Clearly, all the above results can be generalised to the situation where the covariance matrix of the residuals under $A$ is no longer proportional to the identity matrix, but proportional instead to a known positive definite (typically diagonal) matrix.

REFERENCES


