2008

Measurement Error in Auxiliary Information

R. Chambers
University of Wollongong, ray@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

08-08

Measurement Error in Auxiliary Information

Ray Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

MEASUREMENT ERROR IN AUXILIARY INFORMATION

Ray Chambers

Centre for Statistical and Survey Methodology
University of Wollongong

1 Introduction

Survey data are not just the data collected from the responding sample. There are typically many other sources of information about the characteristics of the sampled population that can be used to improve inference. The data contained in these sources is often referred to as auxiliary data in the survey sampling literature. Inference using the survey data that exactly recovers key population characteristics associated with this auxiliary information is said to be calibrated on these characteristics, and is typically viewed as superior to inference that does not necessarily achieve this outcome.

Unfortunately, in most practical situations auxiliary information is not precise. For example, a common situation is where the population mean values of a set of auxiliary variables are assumed known, and survey weights are constructed so that survey estimates of these population means equal their known population values. Weights that are calibrated in this way are used extensively by national statistical agencies. However, it is not unusual that the so-called true values of the population means of the auxiliary variables that are used in construction of the calibrated weights are themselves estimates, perhaps based on administrative records that contain errors, or more often, population means of closely related, but not identical, variables measured by administrative systems. In such cases, the superiority of inference based on weights that are calibrated to incorrect population values is debatable.

This paper considers the impact of such measurement errors in auxiliary information in two somewhat different situations. The first is the calibrated weighting situation described in the preceding paragraph. The second is where marginal population information is available for improving regression estimation, but this marginal information contains errors. In both cases simulation results are presented that demonstrate the impact on inference.

2 Auxiliary information in survey sampling

As pointed out at the start of the previous section, the data obtained from responding sample units (which typically include the variables that are the focus of

inference) are not the only source of statistical data in the context of survey sampling. Other sources include data obtained from all sampled units and data relating to the characteristics of the sampled population. The former are typically useful in allowing us to compare respondents and non-respondents and hence provide an insight into deciding whether the non-response is ignorable, while the latter allow us to compare sampled and non-sampled units in the population and so provide information about whether the sample is representative of the population. In this context, we note that these population data typically constitute summary-type information (e.g. average values) and so permit comparison of corresponding respondent and sample summaries, with a view to perhaps adjusting sample inference in order to recover known population values (i.e. calibration).

In professionally implemented surveys, one typically also has information about the sample design (e.g. method of sampling, stratum boundaries, sample weights and, in some circumstances, clustering information), which is necessary if one needs to take account of the method of sampling in inference. Ideally, these paradata also includes information about the quality of the survey data (e.g. characteristics of measurement error, data editing summaries, interviewer feedback, re-interview information about response error), all of which is useful in shaping our confidence in inference based on the survey data.

As a consequence auxiliary information can be an extremely important component of survey data. It incorporates all the different data sources that provide information about the population from which the sample data were obtained, including information about how the sample was selected and how the non-sampling error is distributed. Ideally, one would like to combine this auxiliary information with the data obtained from the sample units for more efficient inference. In this context, it can be noted that auxiliary information often ensures recognisable samples by allowing comparison of aspects of the joint distribution of auxiliary variables in the respondent sample with corresponding aspects of the joint population distribution. An immediate consequence is that one can then calibrate sample inference so that it recovers these population characteristics. Of course, there remains the issue of deciding just which differences are worth bothering about, and which are not. We will not investigate this important issue here. Instead, we focus on the common situation of calibration to a defined set of population totals.

3 Calibrated survey estimation

A standard situation is where a linear estimator

$$\hat{t}_y^w = \sum_s w_{is} y_i = \mathbf{w}_s^t \mathbf{y}_s \tag{1}$$

of a survey variable Y is required. Here s denotes the n units in sample, and U denotes the population of N units. We use lowercase (uppercase) bold to denote vector (matrix) quantities and a subscript of s to denote sample values. A vector Z of p auxiliary variables is also measured on the sample, with the population values of Y and Z related through the linear regression model

$$\mathbf{y}_U = \mathbf{Z}_U\beta + \mathbf{e}_U. \tag{2}$$

Here a subscript of U denotes a population level quantity. Without loss of generality we shall assume that the population units are ordered so that sample/non-sample decompositions

$$\mathbf{y}_U = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}$$

and

$$\mathbf{Z}_U = \begin{pmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{pmatrix}$$

hold, where the subscript r denotes a quantity defined by the non-sampled population units. As usual, we assume that the residuals in (2) have zero expected values. We also assume that

$$Var(\mathbf{e}_U) = \sigma^2\mathbf{V}_U$$

where

$$\mathbf{V}_U = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$$

is a known positive definite matrix of order N, with sample/non-sample decomposition as shown.

The vector of sample weights used in the estimator (1) is said to be calibrated on Z if this estimator exactly recovers the population totals of the components of this variable. That is, we have

$$\sum\nolimits_s w_{is}\mathbf{z}_i = \mathbf{Z}_s^t\mathbf{w}_s = \sum\nolimits_U \mathbf{z}_i = \mathbf{Z}_U^t\mathbf{1}_N. \tag{3}$$

Here $\mathbf{1}_k$ denotes a vector of ones of dimension k. Suppose now that the method of sampling is non-informative given Z. Then (2) also holds for the sample and it is easy to see that the constraint (3) is equivalent to requiring that the linear estimator (1) is unbiased under this model since then

$$E\left(\hat{t}_y^w - t_y\right) = E\left(\mathbf{w}_s^t\mathbf{y}_s - \mathbf{1}_N^t\mathbf{y}_U\right) = \left(\mathbf{w}_s^t\mathbf{Z}_s - \mathbf{1}_N^t\mathbf{Z}_U\right)\beta = 0.$$

Given the calibration constraint (3), efficient calibrated weights are easily defined. In particular, In the situation of interest, (1) is the best linear unbiased predictor

(BLUP) of the population total of Y provided the weights that define this estimator are of the form (Royall, 1976; Valliant, Dorfman and Royall, 2000, section 2.2)

$$\mathbf{w}_s^{BLUP} = \mathbf{1}_n + \mathbf{H}^t \left(\mathbf{Z}_U^t \mathbf{1}_N - \mathbf{Z}_s^t \mathbf{1}_n \right) + \left(\mathbf{I}_n - \mathbf{H}^t \mathbf{Z}_s^t \right) \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \mathbf{1}_{N-n} \qquad (4)$$

where $\mathbf{I}_n$ denotes the identity matrix of order n and

$$\mathbf{H} = \left(\mathbf{Z}_s^t \mathbf{V}_{ss}^{-1} \mathbf{Z}_s \right)^{-1} \mathbf{Z}_s^t \mathbf{V}_{ss}^{-1}.$$

It is easy to see that the BLUP weights (4) are calibrated on Z since $\mathbf{Z}_s^t \mathbf{H}^t$ is the identity matrix of order p.

The preceding development is essentially how one might motivate calibrated weighting under a model-based approach to sample survey inference. It is not the standard way this idea is developed in the literature, where the model-assisted approach of Deville and Sarndal (1992) is usually followed. Here the idea is to choose the sample weights so that they are calibrated, i.e. they satisfy (3), and are as close as possible to the traditional expansion weights $\mathbf{w}_s^\pi = \left(\pi_i^{-1}; i \in s \right)$ that define the design-unbiased Horvitz-Thompson estimator of the population total of Y. Note that π_i denotes the sample inclusion probability of population unit i. A standard metric for this closeness is

$$Q = (\mathbf{w}_s - \mathbf{w}_s^\pi)^t \Omega_s (\mathbf{w}_s - \mathbf{w}_s^\pi)$$

where Ω_s is a positive definite matrix of order n. Typically this matrix is diagonal, with diagonal element corresponding to sample unit i proportional to that units sample inclusion probability multiplied by the corresponding diagonal element of the covariance matrix $\mathbf{V}_{ss}$. Minimising Q subject to (3) leads to generalised regression (GREG) weights of the form

$$\mathbf{w}_s^{GREG} = \mathbf{w}_s^\pi + \Omega_s^{-1} \mathbf{Z}_s^t \left(\mathbf{Z}_s^t \Omega_s^{-1} \mathbf{Z}_s \right)^{-1} (\mathbf{Z}_U^t \mathbf{1}_N - \mathbf{Z}_s^t \mathbf{w}_s^\pi). \qquad (5)$$

Irrespective of whether (4) or (5) is used to calculate the sample weights in (1), there is a built-in assumption that the imposition of the calibration constraint (3), which ensures unbiasedness under the linear model (2), is a good thing. However, this is not necessarily the case. A situation that occurs reasonably often is where there is another vector-valued variable X, of dimension q, which, when substituted for Z in (2) provides a better fit for Y. However, we do not know the population totals of all the components of X, and so we cannot just calibrate on this alternative auxiliary variable. Note that X and Z can share components (e.g. both could include an intercept term), but there are components of X that are not

in Z. In this situation, we face a dilemma. Do we still proceed with calibration on Z, even though the justification for a linear relationship between Y and Z is weak? Or do we replace Z by X in (4) and (5) and then seek to approximate the unknown population totals associated with X? That is, do we exactly calibrate to a poorly fitting linear model or do we approximately calibrate to a better fitting linear model? In the following section we explore this choice in the context of a realistic business survey example.

4 Using estimated calibration constraints in survey estimation

Consider the following business survey example. Suppose that Y is the total wages paid by a sampled business over a defined period of time, and let X denote the number of employees of the business over the same period. Suppose also that the sampled businesses are drawn from an administrative list (the Business Register), and that for every business on this register we have an approximate value of its size, as defined by the number of its employees at some time in the past. We denote this register size variable by Z. Figure 1 illustrates this situation using some actual business survey data for two sector (G and K) of the register.

The main thing to note about the relationships shown in Figure 1 is that, as one would expect, the linear relationship between Y and X is noticeably stronger than that between Y and Z. In particular, the correlations between Y and X in sectors G and K are 0.9345 and 0.7972 respectively, while those between Y and Z in these sectors are 0.9108 and 0.6293. As a consequence, our preference is to model Y linearly in X, rather than to model Y linearly in Z. That is, our preferred model is

$$E(Y|X) = \alpha_X + \beta_X X \tag{6}$$

and

$$Var(Y|X) = \sigma_X^2 X^2$$

rather than

$$E(Y|Z) = \alpha_Z + \beta_Z X \tag{7}$$

and

$$Var(Y|Z) = \sigma_Z^2 Z^2.$$

Note that the assumption of quadratic heteroskedasticity in both (6) and (7) relates to the fact that under logarithmic transformation of Y, X and Z, the plots in Figure 1 become quite linear, with homoskedastic errors.

Unfortunately, although we prefer (6), estimation of the population total of Y

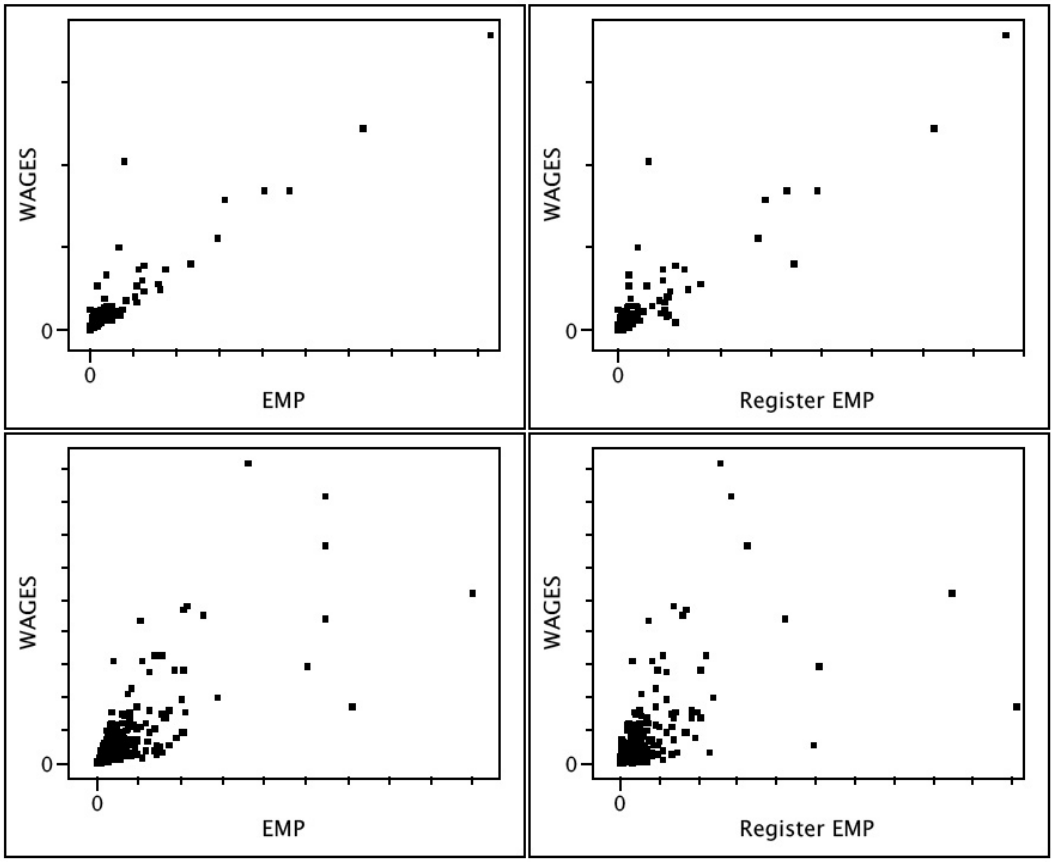

Figure 1: Total wages paid (Y = WAGES) vs. actual employment (X = EMP) and vs. register employment (Z = Register EMP) for two groups of sampled businesses in a business survey. The top row corresponds to 768 businesses in sector G, while the bottom row corresponds to 1005 businesses in sector K.

under this model using either the BLUP weights (4) or the GREG weights (5) requires specification of the sector totals of X, which are unknown. In contrast, we can easily extract the sector totals of Z from the register and hence calculate (4) or (5) when (7) is the model of choice. This leaves us with three options on how to proceed.

Option 1. Use (7) to construct the survey weights. We refer to this as the *Alt*(ernate) option.
Option 2. The correlation between X and Z in the sample (0.9836 in Sector G and 0.9096 in Sector K) is larger than the correlation of either of these variables with Y. Consequently, we use (6) to construct the survey weights assuming that the sector totals of X and Z are the same. That is, we replace Z by X in (4) and (5), and replace the unknown sector totals of X by the known sector totals of Z. We refer to this as the *Sub*(stitution) option.
Option 3. As in *Option* 2, we use (6) to construct the survey weights, but now we estimate the sector totals of X using Z, and then use these estimated sector totals in (4) and (5). That is, we calibrate to estimated population totals under (6). We refer to this as the *Pred*(iction) option. Note that individual values of X and Z are highly heteroskedastic, so we use a simple outlier robust method to predict the unknown population (i.e. sector) total t_X of X:

$$\hat{t}_X = \underset{i \in s}{median} \left\{ \frac{x_i}{z_i} \right\} \times t_Z.$$

In order to evaluate these three options, we carried out a small simulation study based on sub-sampling the data in Figure 1. The study used two distinct sample designs, independently repeated 1000 times, to draw the samples from these populations, and was applied separately in each sector. The first of these is denoted STRS. It had a sample size of $n = 50$ in each sector, with four Z-based strata in each sector and stratum boundaries defined so that the Z totals in the strata were approximately equal. The sample allocations to these strata were 13, 13, 12, 12 (sector G) and 15, 15, 15, 5 (sector K), with the top (fourth) stratum in each case completely enumerated, and with independent simple random samples taken without replacement from the remaining three strata. The second sample design is denoted PPZ. This still had a sample size of $n = 50$ and a completely enumerated top stratum, but the remaining sample units were selected using a probability proportional to size (as measured by Z) sampling scheme.

Table 1 shows the values of relative bias and relative RMSE (both expressed as percentages) that were obtained in the study. Note that two sets of results are shown for the STRS sampling scheme. The first is for estimation (BLUP and GREG) based on the assumption that the underlying linear model holds across

Sector		Bias				RMSE			
		Pref	*Alt*	*Sub*	*Pred*	*Pref*	*Alt*	*Sub*	*Pred*
G	STRS1	0.97	0.66	-13.32	0.16	13.17	12.76	16.69	13.48
		-4.00	-5.98	-14.53	-4.63	9.22	12.00	16.00	10.10
	STRS2	1.11	0.70	-9.12	-0.14	12.24	11.24	13.72	11.72
	PPZ	0.28	-0.15	-11.71	-0.15	11.64	9.60	14.36	11.43
		-4.77	-6.57	-14.61	-4.98	8.12	13.12	15.82	9.66
K	STRS1	0.44	0.24	-5.91	-0.43	11.32	13.29	12.19	11.93
		-5.69	-5.75	-11.50	-6.48	10.03	11.81	13.83	11.27
	STRS2	1.07	-0.76	-7.21	-0.85	12.02	14.00	12.49	12.45
	PPZ	0.16	0.05	-6.34	-1.73	9.98	11.32	11.40	11.44
		-3.86	-2.64	-9.68	-5.57	8.42	11.23	11.96	10.57

Table 1: Values of relative bias and relative RMSE (in percent) generated in the simulation study. Note that the top entry in the STRS1 and PPZ rows is for the BLUP and the bottom entry is for the GREG.

the entire sector. For the *Pred* option this requires that we estimate the sector level total of X. This is denoted STRS1 in Table 1. The second is where this model holds at size stratum level, in which case BLUP and GREG coincide. This is denoted STRS2 in Table 1. Here implementation of the *Pred* option requires that we estimate stratum level totals for X. Also, we provide results when (6) is assumed to hold and both stratum and sector level totals for X are known. Although this option is not really available to us, it is the one that we actually prefer, and so provides a benchmark against which we can compare the other estimation options. It is denoted *Pref* in Table 1.

Inspection of the results in Table 1 confirms that *Pref* is the best of the four estimation methods considered in the study. However, as already noted, this estimation method cannot be implemented because it relies on the availability of sector and stratum level totals for X. In contrast, *Sub* is not a good estimation method, exhibiting bias in all situations examined in the simulation study. Consequently, the choice is between *Alt* and *Pred*. Both estimation methods exhibit similar biases, and neither appears superior in terms of RMSE performance.

In order to throw some more light on the differences between *Alt* and *Pred*, we also compared them in terms of their revision error, defined as their relative difference from *Pref*. The reason for this is simple in many situations population values of X do eventually become available as administrative systems are updated. In this case we want our original estimates to require as little revision as possible when the updated auxiliary information is brought on line. In Figure

2 we present boxplots showing the distribution of these relative differences under both versions of the STRS sampling strategy. The large (and stable) bias in *Sub* is evident. However, we now see that *Pred* is much better than *Alt* in terms of smaller revision variability.

In summary therefore we note that although *Alt* is the optimal under (7), the better fit of (6) means that *Pred* can be more efficient, and is certainly preferable in terms of smaller revision error as measured by variability in differences from *Pref*. However, one needs to be careful here, because the performance of *Pred* depends very much on how the sector and stratum totals of X are estimated. In our case we used an outlier robust methodology for this purpose. Less robust methods can substantially degrade the performance of *Pred*. Also, one has to take account of the fact that variance estimation for *Pred* is more complicated than that for *Alt*. The latter is a straightforward regression estimator, and well-known methods of variance estimation can be used. In contrast variance estimation for *Pred* is a special case of two phase variance estimation, and so its variance estimator contains two terms, one being the usual variance estimator assuming the estimated sector and stratum totals for X are correct, and the other generated by the variances of the estimators of these quantities.

5 Beyond calibration - integrating auxiliary information into parametric inference

So far in this paper, we have focussed on a traditional inference target for a sample survey, i.e. the population total of Y. We now turn to a more analytic use of the survey data, where the target of inference is a parameter of the stochastic process that is assumed to have generated the population values of Y. In particular, we assume that the population vector $\mathbf{y}_U$ is generated as a random draw realisation of a random vector $\mathbf{Y}_U$ with density f_Y, which is known up to the value of a parameter θ. Our aim in this situation is to use the survey data to calculate the maximum likelihood estimate of θ. The approach we take is based on application of the Missing Information Principle (Orchard and Woodbury, 1972) within the inferential framework described in Breckling et al. (1994). See also Chambers and Skinner (2003, Chapter 2).

As in the previous development, our first step is to identify the available survey data, which we denote by $\mathbf{D}_s$. Note that this consists of all the data sources described in section 2, and, assuming full response (or ignorable non-response), contains in particular the sample values $\mathbf{y}_s$ of Y, the population vector $\mathbf{i}_U$ of sample inclusion indicators and the population matrix $\mathbf{Z}_U$ of values of the auxiliary variables. It can also contain summary information for Y, either as measured at

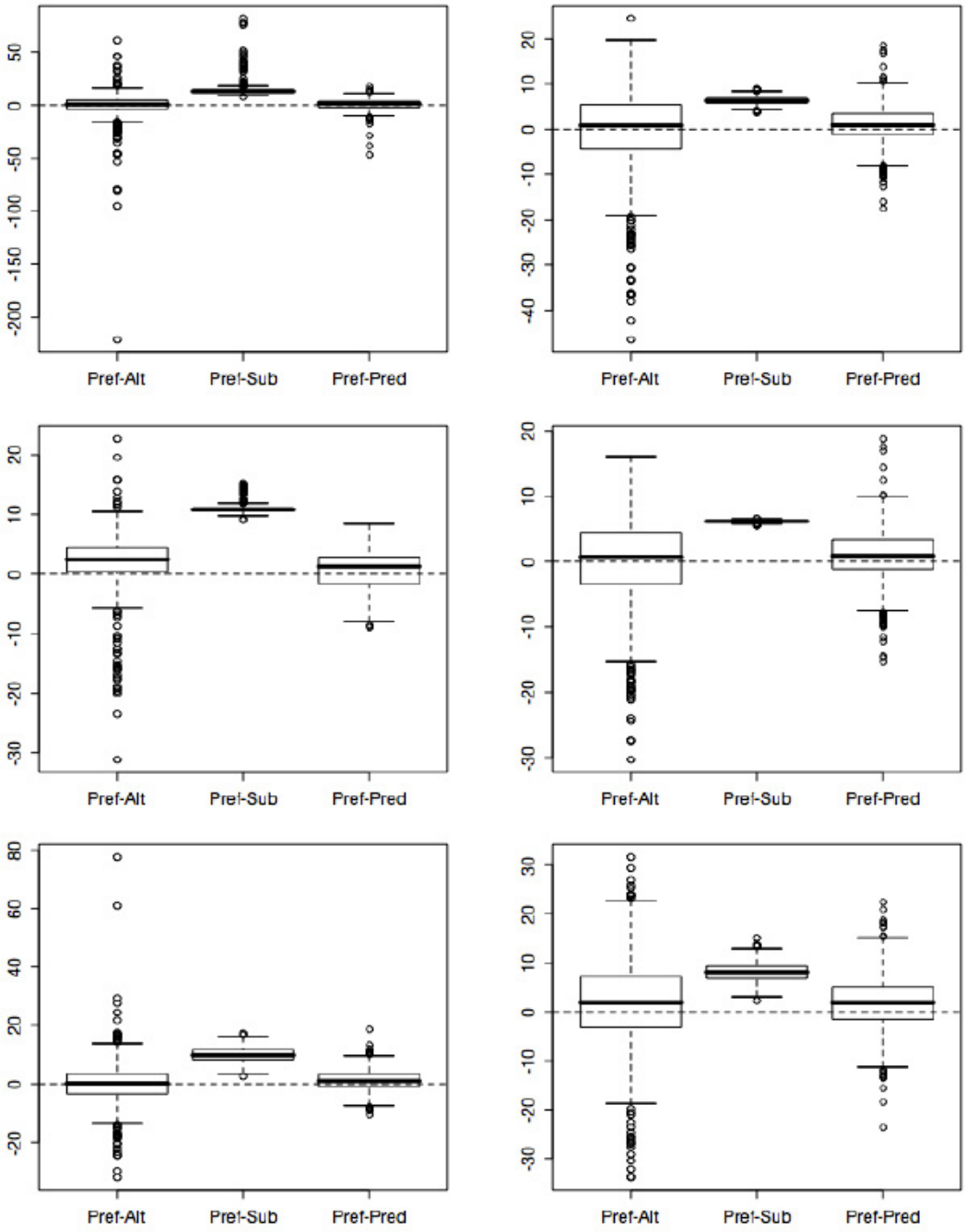

Figure 2: Boxplots showing relative differences (in percent) between *Pref* estimates and estimates based on *Alt*, *Sub* and *Pred* options. These are denote Pref-Alt, Pref-Sub and Pref-Pred respectively. Top row is STRS1 with GREG, middle row is STRS1 with BLUP and bottom row is STRS2 with BLUP, which is the same as GREG. Left plot is sector G and right plot is sector K.

the time of the survey, or, more likely, at some time in the past. Corresponding to $\mathbf{D}_s$, we can then define (at least conceptually) its population equivalent $\mathbf{D}_U$, which will have a distribution $f_D(\mathbf{D}_U; \Theta)$ that is parameterised by Θ. Note that θ is then a component of Θ or θ can be obtained by a one to one transformation of components of Θ. In either case, if we can calculate the maximum likelihood estimate of Θ, we can then derive the corresponding estimate of θ. Consequently we now focus on maximum likelihood estimation of Θ.

To start, we note that maximum likelihood analysis of Θ depends on two key concepts. The first is the score function, which is the derivative of the logarithm of the likelihood function for Θ. In particular, the ML estimator (MLE) is the value of Θ where the score function is zero. The second is the information function for Θ, which is negative of the derivative of the score function with respect to Θ. The estimated variance of the MLE is the inverse of the value of the information function at the MLE.

Breckling *et al.* (1994) develop two key identities that can be used to compute the score and information functions. The first states that the score $sc_s(\Theta)$ for Θ generated by the survey data $\mathbf{D}_s$ is the conditional expectation, given these data, of the score $sc_U(\Theta)$ for Θ generated by the corresponding population data $\mathbf{D}_U$, i.e.

$$sc_s(\Theta) = E\left\{sc_U(\Theta) \mid \mathbf{D}_s\right\}. \tag{8}$$

The second states that the information $info_s(\Theta)$ for Θ generated by the survey data $\mathbf{D}_s$ is the conditional expectation, given these data, of the information $info_U(\Theta)$ for Θ generated by the population data $\mathbf{D}_U$ minus the corresponding conditional variance of the population score, i.e.

$$info_s(\Theta) = E\left\{info_U(\Theta) \mid \mathbf{D}_s\right\} - Var\left\{sc_U(\Theta) \mid \mathbf{D}_s\right\}. \tag{9}$$

A widely used alternative to the maximum likelihood approach outlined above is to estimate θ by maximising its pseudo-likelihood (Pfeffermann, 1993). This is a model-assisted approach that can be motivated as follows. Recollect that our inference is based on a model where $\mathbf{Y}_U \sim f_Y(\mathbf{y}_U; \theta)$. Consequently, if $\mathbf{y}_U$ were observed, θ would be estimated by solution of

$$sc_U(\theta) = \frac{\partial \log f_Y(\mathbf{y}_U; \theta)}{\partial \theta} = 0.$$

Now, for any specified value of θ, $sc_U(\theta)$ is a well-defined function of the values in $\mathbf{y}_U$ (it is the so-called 'census value' of this score function), and so can be estimated using standard survey sampling methods. In particular, if the population

values of Y are independent, with $f_i(y;\theta)$ denoting the density of Y for population unit i, then

$$sc_U(\theta) = \sum_U \frac{\partial \log f_i(y_i;\theta)}{\partial \theta} = \sum_U sc_i(\theta).$$

Given a set of survey weights $(w_{is}; i \in s)$, this marginal score function can be estimated by

$$sc_w(\theta) = \sum_s w_{is} sc_i(\theta) \tag{10}$$

and the maximum pseudo-likelihood estimator of θ is then the solution to $sc_w(\theta) = 0$. Note that there is no concept that is equivalent to an information function under this approach, with the estimated variance of the maximum pseudo-likelihood estimator computed using standard survey sampling sandwich type methods (Binder, 1983). Also, in this context there is no agreed mechanism for including auxiliary information into inference. However, an obvious way to do this is by suitable calibration of the survey weights used in (10).

6 Efficient linear regression given marginal population information

We now apply the likelihood-based theory described in the previous section in the context of a particular case of imprecise auxiliary information. Suppose that the sample survey measures the values y_i and x_i of two scalar variables, Y and X respectively, for a sample s of n units from a population U of N units. Our aim is to use the survey data to estimate the parameters α, β and σ^2 that characterise the popular population-level linear regression model

$$y_i = \alpha + \beta x_i + \sigma e_i \tag{11}$$

where the errors e_i are assumed to be independently and identically distributed standard Gaussian random variables. We also assume that Y is independent of the (random) sample inclusion indicator I given X. That is, the sampling method is non-informative for the parameters of (11).

It is easy to see that in the absence of any other information the maximum likelihood estimators for α, β and σ^2 are then the usual sample-based MLEs:

$$\hat{\beta} = \Big(\sum_s x_i(x_i - \bar{x}_s)\Big)^{-1} \sum_s x_i(y_i - \bar{y}_s) \tag{12}$$

$$\hat{\alpha} = \bar{y}_s - \hat{\beta}\bar{x}_s \tag{13}$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_s (y_i - \hat{\alpha} - \hat{\beta} x_i)^2. \tag{14}$$

However, now suppose that we have extra information In particular, suppose that we know the population means $\bar{y}_U$ and $\bar{x}_U$ of Y and X. This type of marginal information may be available from separate population registers, for example. Given the sample values of Y, this immediately yields the average value $\bar{y}_r$ of the non-sampled units in the population. Now the estimators (12), (13) and (14) are no longer the MLEs of the parameters of (11). In order to derive the form of the MLEs in this situation we use the approach described in the previous section. Put $\theta = (\alpha, \beta, \sigma^2)$. The components of the population level score function for this parameter in this case are

$$sc_1(\theta) = \sigma^{-2} \sum_U (y_i - \alpha - \beta x_i)$$

$$sc_2(\theta) = \sigma^{-2} \sum_U x_i(y_i - \alpha - \beta x_i)$$

and

$$sc_3(\theta) = -N/2\sigma^2 + \sum_U (y_i - \alpha - \beta x_i)^2/2\sigma^4.$$

Suppose individual population values of X known. From (8), we see that the components of the full information sample score function (where the subscript of s denotes conditioning on the survey data) are now

$$sc_{1s}(\theta) = \sigma^{-2} \sum_U (E_s(y_i) - \alpha - \beta x_i)$$

$$sc_{2s}(\theta) = \sigma^{-2} \sum_U x_i(E_s(y_i) - \alpha - \beta x_i)$$

and

$$sc_{3s}(\theta) = \frac{1}{2\sigma^2}\left[-N + \frac{1}{\sigma^2}\left\{\sum_U (E_s(y_i) - \alpha - \beta x_i)^2 + \sum_U Var_s(y_i)\right\}\right].$$

In order to evaluate the conditional expectations and variances in these expressions, observe that for non-sampled unit i

$$\begin{pmatrix} y_i \\ \bar{y}_r \end{pmatrix} \mid \mathbf{x}_U \sim N\left[\begin{pmatrix} \alpha + \beta x_i \\ \alpha + \beta \bar{x}_r \end{pmatrix}, \begin{bmatrix} \sigma^2 & (N-n)^{-1}\sigma^2 \\ (N-n)^{-1}\sigma^2 & (N-n)^{-1}\sigma^2 \end{bmatrix}\right].$$

It immediately follows that

$$y_i \mid \mathbf{x}_U, \bar{y}_r \sim N\left[\bar{y}_r + \beta(x_i - \bar{x}_r), \sigma^2\left\{1 - (N-n)^{-1}\right\}\right]$$

which in turn allows us to write

$$sc_{1s}(\theta) = \sigma^{-2}\left[\sum_s (y_i - \alpha - \beta x_i) + (N-n)(\bar{y}_r - \alpha - \beta \bar{x}_r)\right] \qquad (15)$$

$$sc_{2s}(\theta) = \sigma^{-2}\left[\sum_s x_i(y_i - \alpha - \beta x_i) + (N-n)\bar{x}_r(\bar{y}_r - \alpha - \beta \bar{x}_r)\right] \qquad (16)$$

and

$$sc_{3s}(\theta) = \frac{1}{2\sigma^2}\left[-(n+1) + \frac{1}{\sigma^2}\left\{ \begin{array}{l} \sum_s (y_i - \alpha - \beta x_i)^2 \\ +(N-n)(\bar{y}_r - \alpha - \beta \bar{x}_r)^2. \end{array} \right\}\right] \tag{17}$$

Setting (15), (16) and (17) to zero and solving for α, β and σ^2 leads to the full information MLEs (FIMLEs) for these components:

$$\hat{\beta}^{FIMLE} = \frac{\sum_s \{x_i(y_i - \bar{y}_s)\} + n\bar{x}_s(\bar{y}_s - \bar{y}_U) + (N-n)\bar{x}_r(\bar{y}_r - \bar{y}_U)}{\sum_s \{x_i(x_i - \bar{x}_s)\} + n\bar{x}_s(\bar{x}_s - \bar{x}_U) + (N-n)\bar{x}_r(\bar{x}_r - \bar{x}_U)} \tag{18}$$

$$\hat{\alpha}^{FIMLE} = \bar{y}_U - \hat{\beta}^{FIMLE}\bar{x}_U \tag{19}$$

and

$$\begin{array}{c} \hat{\sigma}^{2,FIMLE} = (n+1)^{-1}\sum_s (y_i - \hat{\alpha}^{FIMLE} - \hat{\beta}^{FIMLE}x_i)^2 \\ +(N-n)(\bar{y}_r - \hat{\alpha}^{FIMLE} - \hat{\beta}^{FIMLE}\bar{x}_r)^2. \end{array} \tag{20}$$

It is interesting to note that (18), (19) and (20) are identical to weighted least squares estimators of these parameters defined by an extended sample consisting of the data values for the units in s (each with weight equal to one) plus an additional data value (with weight equal to Nn) corresponding to the known non-sample means $\bar{y}_r$ and $\bar{x}_r$.

An alternative to the above MLEs is to use (GREG-based) calibrated weighting combined with maximum pseudo-likelihood to estimate the parameters of interest. In this context we note that there are three calibration constraints - the population size N, the population mean of X and the population mean of Y. Using (5), the corresponding calibrated weights are given by

$$\mathbf{w}_s^{GREG} = \left(w_{is}^{GREG}; i \in s\right) = \mathbf{w}_s^{\pi} + \mathbf{D}_s^{\pi}\mathbf{Z}_s\left(\mathbf{Z}_s^t\mathbf{D}_s^{\pi}\mathbf{Z}_s\right)^{-1}\left(\begin{array}{c} 0 \\ N\bar{y}_U - \sum_s \pi_i^{-1} y_i \\ N\bar{x}_U - \sum_s \pi_i^{-1} x_i \end{array} \right)$$

where $\mathbf{D}_s^{\pi} = diag\left(\mathbf{w}_s^{\pi}\right)$ and $\mathbf{Z}_s = [\mathbf{1}_n\ \mathbf{y}_s\ \mathbf{x}_s]$. The maximum pseudo-likelihood estimators based on these weights are easily seen to be

$$\hat{\beta}^{GREG} = \left\{\sum_s w_{is}^{GREG} x_i(x_i - \bar{x}_s^{GREG})\right\}^{-1}\sum_s w_{is}^{GREG} x_i(y_i - \bar{y}_s^{GREG}) \tag{21}$$

$$\hat{\alpha}^{GREG} = \bar{y}_s^{GREG} - \hat{\beta}^{GREG}\bar{x}_s^{GREG} \tag{22}$$

and

$$\hat{\sigma}^{2,GREG} = N^{-1}\sum_s w_{is}^{GREG}(y_i - \hat{\alpha}^{GREG} - \hat{\beta}^{GREG}x_i)^2 \tag{23}$$

where

$$\bar{y}_s^{GREG} = \left\{\sum_s w_{is}^{GREG}\right\}^{-1}\sum_s w_{is}^{GREG} y_i$$

and $\bar{x}_s^{GREG}$ is defined similarly. These estimators are denoted as GREG estimators in what follows.

How do FIMLE and GREG compare in terms of efficiency? And, in particular, how do they compare when the population averages of Y and X are measured with error? In order to answer these questions we again carried out a small simulation study, consisting of 1000 independent simulations of a set of population values. A sample was then selected from each population according to a specified sampling design and the resulting values of (18), (19) and (20) and (21), (22) and (23) calculated. The population data used in these simulations were generated according to

$$y_i = 5 + x_i + e_i$$

with values of X independently drawn from the standard lognormal distribution and the regression errors e_i independently generated as standard Gaussian. Two sampling methods were investigated for a variety of sample and population sizes. The first was simple random sampling without replacement (SRS, $\pi_i = nN^{-1}$), with the second corresponding to probability proportional to Y sampling without replacement (PPY, $\pi_i = nN^{-1}y_i\bar{y}_U^{-1}$). Note that this second sampling method is informative.

In order to evaluate performance when calibration constraints are approximate, rather than exact, we considered three scenarios. The first (scenario A) consisted of perfect information, i.e. $\bar{y}_U$ and $\bar{x}_U$ are known. The second (scenario B) emulated a situation where these constraints are estimated with census level error. This would be the case, for example, if the values $\hat{\bar{y}}_U$ and $\hat{\bar{x}}_U$ used as calibration constraints are the population averages of variables that have the same expected values as Y and X, so

$$\hat{\bar{y}}_U = \bar{y}_U + N^{-1/2}z_{Yi}$$

and

$$\hat{\bar{x}}_U = \bar{x}_U + N^{-1/2}z_{Xi}$$

where z_{Yi} and z_{Xi} are independent standard Gaussian variables. Finally, the third (scenario C) related to a situation where the constraints are unbiased estimates based on a larger survey with a sampling fraction of 20 percent, i.e.

$$\hat{\bar{y}}_U = \bar{y}_U + (N/5)^{-1/2}\, z_{Yi}$$

and

$$\hat{\bar{x}}_U = \bar{x}_U + (N/5)^{-1/2}\, z_{Xi}.$$

Table 2 shows the relative efficiencies of the FIMLE and GREG estimators in the study. Note that these efficiencies are defined as the ratio of the 5% trimmed

RMSE to the corresponding 5% trimmed RMSE of the simple sample-based estimators (12), (13) and (14) in the case of the SRS sample design, and as the ratio of the 5% trimmed RMSE to the corresponding 5% trimmed RMSE of the simple inverse pi-weighted maximum pseudo-likelihood estimator in the case of the PPY sample design. Trimmed RMSEs were used in all cases to exclude some extreme values that were generated (in particular, by GREG) in the simulation study. Inspection of Table 2 confirms what one expects. It is the FIMLE of α that benefits most from the inclusion of accurate marginal population information. However there are also non-negligible gains for estimation of β and σ^2. Furthermore, these gains increase when a highly informative sampling method (PPY) is used. As the size of the errors in the auxiliary information increases, however, these gains vanish, and, in many cases, turn into losses. In particular, if the auxiliary information has the same margin of error as an estimate derived from a large survey, there are essentially no gains from including it in inference. Turning to GREG, we see that the case for including the extra marginal information in calibration is much weaker than for FIMLE. In particular, even with accurate auxiliary information it is only the GREG of α that is improved relative to ignoring this information, and these gains quickly evaporate with increasing error in the auxiliary information. Overall, our results provide evidence that using FIMLE to integrate auxiliary information into inference can be useful. However, integrating this information via GREG-type weight calibration did not really work in the situations that we investigated.

7 Summary and conclusions

In this paper we investigate two important areas where auxiliary data, if accurate, can improve sample survey inference. However, such data often contain errors, and we use two illustrative simulation studies to explore the impact of these errors on inference. Not surprisingly, the potential gains from use of these data can quickly disappear if they contain errors. However, there are strategies that one can adopt to ensure that inference remains (relatively) robust in the presence of such errors. Thus, it can pay to explore methods of predicting what the correct population values might be in the case where these values are missing for auxiliary variables in a model that fits the sample data well. Conversely, some methods of including auxiliary information in inference (e.g. FIMLE) can be robust to small errors in auxiliary data, while others (e.g. GREG) can be quite sensitive.

Finally, it is appropriate to comment that in the case of FIMLE, the Missing Information Principle provides a mechanism for including uncertainty about the auxiliary data in inference, provided that one is willing to jointly model variability

Parameter	Scenario	$N = 500, n = 20$	$N = 1000, n = 50$	$N = 5000, n = 200$
			SRS	
α	A	133.97	144.75	149.52
		102.97	127.36	143.01
	B	115.50	111.06	115.90
		83.85	100.84	112.33
	C	85.66	79.57	78.09
		63.51	70.76	75.09
β	A	105.89	101.95	100.54
		81.29	89.71	96.11
	B	103.68	99.68	99.60
		73.43	88.96	96.39
	C	99.85	95.37	100.47
		71.12	83.77	92.62
σ^2	A	102.34	100.12	100.08
		84.00	93.54	99.43
	B	103.21	100.52	100.32
		77.87	87.57	96.69
	C	98.82	93.73	99.12
		63.31	76.91	93.61
			PPY	
α	A	200.88	210.23	221.53
		118.46	143.12	158.57
	B	136.37	138.76	152.02
		98.35	120.24	134.79
	C	84.31	75.58	81.76
		69.21	74.37	89.27
β	A	109.14	110.25	116.96
		62.77	73.07	80.85
	B	107.27	111.74	120.81
		64.74	69.79	76.89
	C	103.08	107.21	116.53
		54.11	57.12	65.93
σ^2	A	105.94	105.55	111.14
		78.30	89.06	91.48
	B	107.98	106.99	109.03
		77.00	82.26	90.04
	C	98.88	100.96	102.44
		61.58	71.09	87.09

Table 2: Relative efficiencies (in percent) obtained in the simulation study. Note that the top entry in each row is for FIMLE and the bottom entry is for GREG.

of these data and of the response variable of interest. This type of extended FIMLE is not developed in this paper, and remains a subject for further research.

8 References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, **62**, 349-363.

Chambers, R.L. and Skinner, C.J. Eds. (2003). *Analysis of Survey Data*. Chichester: John Wiley.

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Orchard, T. and Woodbury, M.A. (1972). A missing information principle: theory and application. *Proc. 6th Berkeley Symp. Math. Statist.*, **1**, 697-715.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, **61**, 317-337.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657 - 664.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley.