IMPROVED GLS ESTIMATION IN NCES SURVEYS

Steven Kaufman, National Center of Education Statistics: Bonnie Li, Synectics: Fritz Scheuren, The George Washington University
Bonnie Li, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd. #305, Arlington VA 22201

KEY WORDS: Generalized Least Squares, Winsoring Weights, Reducing Mean Square Errors

For the first time, in 1993-1994, the private school component of the Schools and Staffing Survey (SASS) and the Private School Survey (PSS) are being fielded in the same year. Even though these two surveys measure some of the same variables, the results between the surveys will not agree.

PSS and SASS both measure numbers of schools, numbers of teachers, and numbers of students. Conventional simple or raking ratio adjustment procedures could be used to adjust sample weights so that the SASS estimates agreed with the much larger PSS for each of the three totals separately. Such approaches do not work, though, if the weights are to be adjusted so that **all** three SASS estimates agree simultaneously.

As we reported at last year's meetings (Holt et al., 1994), Generalized Least Squares (GLS) techniques is an alternative that offers promise. While the asymptotic properties of GLS and GLS-like estimators are attractive, their finite sampling properties are not necessary desirable. To avoid some of the operational concerns with GLS procedures found in the 1990-1991 experiment, our plan for the 1993-1994 surveys is to follow a three-step process:

1. For the largest schools, GLS reweighting will not be carried out; instead, a direct use of the PSS cases is to be attempted where, through statistical matching of SASS with PSS, the SASS data will be added to one or more of the PSS observations in what is called a "mass imputation" procedure.

2. To improve further the adjustment process, a multivariate ratio adjustment (like in Olkin, 1958) is to be made within moderately sized domains of SASS -- before the GLS procedure is undertaken.

3. Only then will the resulting new SASS weights be carried forward to a GLS estimation step along the lines described in the next section.

Our expectations of these modified procedures were both that they would lead to improvements in SASS mean square error <u>and</u> that operational difficulties would be lessened. The partial results obtained so far bear this out.

Generalized Least Squares (GLS)

For NCES Private School Surveys, the Generalized Least Squares (GLS) techniques advocated by Deville and Sărndal (1992) can be used, as in Imbens and Hellerstein (1993).

To discuss the basic algorithm employed in Generalized Least Squares, it is necessary to define some notation; in particular --

w_i is the original SASS Private School base weight for the ith SASS observation, $i=1,\ldots,n$.

t_i is the SASS total of teachers for the ith SASS observation, $i=1,\ldots,n$.

s_i is the SASS total of students for the ith SASS observation, $i=1,\ldots,n$.

N is the total estimated number of schools, as given by PSS.

T is the total estimated number of teachers, as given by PSS.

S is the estimated total number of students, as given by PSS.

In reweighting SASS, three constraints are imposed on the new weights u_i,

$$\sum u_i = N$$

$$\sum u_i t_i = T$$

$$\sum u_i s_i = S$$

For our application, the new weights u_i, subject to these constraints, are to be chosen (as in Burton 1989) to minimize a loss function that can be written as the sum of squares

$$\sum (u_i - w_i)^2$$

This is perhaps the simplest and most straightforward loss function that might be chosen. Motivating it here is outside our present scope; except to say that the sensitivity of the results to the loss function chosen (e.g., Deville and Särndal, 1992 and Deville et al., 1993) seems not to be too great (but this is, in part, an application issue and will be among the areas for future study). Anyway, the usual Lagrange multiplier formulation of this problem yields after some algebra that the new weights are of the form

$$u_i = w_i + \lambda_1 + \lambda_2 t_i + \lambda_3 s_i$$

where the λ_i are obtained from the matrix expression

$$\underline{d} = M\underline{\lambda}$$

with the vector $\underline{d}$ consisting of three elements, each a difference between the corresponding PSS and SASS totals for schools (first component), teachers (second component), and students (third component); in particular

$$N - \sum w_i$$

$$T - \sum w_i t_i$$

$$S - \sum w_i s_i$$

where the summations are over the SASS sample observations and the quantities: N, T, and S are known PSS totals for schools (N), teachers (T), and students (S) respectively. The matrix M is given by

$$\begin{matrix} n & \sum t_i & \sum s_i \\ \sum t_i & \sum t^2_i & \sum t_i s_i \\ \sum s_i & \sum t_i s_i & \sum s^2_i \end{matrix}$$

and $\underline{\lambda}$ is the vector of unknown GLS adjustment factors obtained from

$$\underline{d} = M\underline{\lambda}$$

Notice that the M matrix is based solely on the unweighted sample relationships among schools, teachers, and students. This is not an essential feature of our approach; a weighted version of the M matrix could have been used -- with, of course, a corresponding change in the loss function to be minimized.

Olkin-like multivariate ratio estimation

An old idea of Olkin(1958) forms a starting point here. Assume we have a total τ, to be estimated from a sample. Olkin proposed a multivariate ratio estimator of the form Y composed of a sum

$$Y = \sum a_i R_i X_i$$

where the a_i are positive and add to 1, the X_i are known outside totals and the R_i are conventional ratios estimated from the sample of τ and X_i.

How cast the Olkin procedure in the PSS and SASS setting? The multivariate ratio Olkin proposed could, in principle, consist of any number of ratio estimates being added together and averaged in some way by the a_i. Note that in our application there are only three outside totals: X_1 for schools, X_2 teachers and X_3 students -- so the expression has been simplified for this analysis.

For this paper, the a_i are simply chosen to be equal to one-third; however, a more natural approach would be to select them so as to minimize the variance of Y. Given the complex sample design of SASS, though, this has been left for the future.

In principle, an Olkin adjustment to the original weights could be produced within whatever domain is desired; then in order to determine the "new" weight for that domain, all the cases would be adjusted such that they would have new weights

$$u_i = Rw_i$$

where the overall ratio R is obtained by taking Y and dividing it by the corresponding estimate obtained from the original sample.

The intuition is that if the Olkin estimation was carried out for small (appropriate) subdomains, then there would be a direct benefit from this step in those subdomains. Further, the overall PSS/SASS differences would shrink appreciably, minimizing any harm that GLS might do. To try something to check these intuitions, it might be enough to use our greatly simplified Olkin-like approach over suitable subdomains (leaving for later, as already mentioned, a way to choose the a_i to minimize the variance of the estimator).

PSS and SASS Data for 1993-94

As noted earlier, it seems natural to use the PSS figures for schools, students, and teachers as the standard and to adjust the SASS estimates correspondingly; that is what we have done here. To fix ideas and to simplify our discussion, only private Nonsectarian Regular Schools will be examined. There were two basic steps taken which are listed below.

1. Based on an initial visual inspection, we identified about a dozen large schools for which some form of mass imputation, rather than reweighting might be the adjustment of choice.

2. With the remaining SASS sample schools and the remaining universe of PSS schools (See figure 1 below), we then calculated Olkin-like factors by school size to begin the adjustment process.

Figure 1. -- PSS and SASS originally weighted estimates compared

	SASS	PSS
Schools	2524	2186
Teachers	52868	49587
Students	514569	463263

3. Then, a GLS adjustment followed, using the adjustment formula, shown below of

$$u_i = w_i - 0.758 + 0.04006t_i - 0.0032s_i$$

The large negative value for $\lambda_1 = -0.758$ meant that for small schools (with only a few teachers) the possibility of very small weights existed. Similarly, for large schools (with many students) the possibility of negative weights existed (since λ_3 is negative too). Our examination of the weights showed, in fact, that three were negative and three very small.

Another look at data plots identified these schools as cases that were away from the basic scatter -- so we excluded them as well. Another 20 or so small schools had weights (between about 0.2 and 0.7). For these schools, we employed a simple winsorizing routine (and added +0.5) -- so that when subjected again to the GLS algorithm they would not be unacceptably small.

Redoing the Olkin and GLS steps with this slightly smaller set of SASS sample cases yielded an acceptable result -- no negative cases and none that were judged to be too small.

Evaluation of Adjustments

There are two ways we will evaluate our results. Each represents an alternate course of action:

1. One possible course of action might be to do nothing. Here we will compare our method to the original SASS weighted results.

2. Another course of action might be to carry out a simple GLS adjustment, without also introducing Olkin-like factors. Here we will be comparing the Olkin-GLS weights (and estimates) with what would have happened if only a GLS estimate had been attempted.

To be consistent with what has been done already, we look only at the SASS sample cases that were finally subjected to an Olkin GLS weight adjustment. Figure 2 displays the original, unadjusted GLS and Olkin GLS weights in the form of a scatterplot matrix. As can be seen, for these cases

- Visually, the three sets of weights appear close; however, at the bottom of the standard GLS weight distribution, there are about 30+ negative weights.

- Notice also that the regression means differ overall too. The standard GLS mean is closest to the original, since it does not adjust the weights separately by school size; also the fit between the Olkin GLS is somewhat poorer than is true for the unadjusted GLS, again for the same reason.

The real test of the methods is how close they come to improving not only overall totals but also the totals by school size. To examine this, a comparison was made for SASS schools, teachers, and students by school size as a percent of the corresponding PSS total. While not uniformly better, the Olkin GLS method demonstrated considerable superiority, suggesting we are on the right track.

Conclusions and Areas for Future Study

The work done so far on intersurvey consistency is gratifying in that a clear improvement has been obtained. There are many issues to face, though, as we try to learn more. Among these are

- Can we find a better, more systematic way of handling outliers (e.g., negative and small weights) ahead of time?
- Using mass imputation is only mentioned in the paper. How would that work in these two NCES surveys?
- Can we unite the Olkin and GLS techniques into a single adjustment (as the theory seems to suggest)?
- What about integrating still other information from PSS into SASS (say, information on Community type)? Via a raking version, perhaps?
- Is there a way to calculate variances for an Olkin GLS estimator that is not any more computationally intensive than for the current SASS estimator?
- What about other GLS loss functions? Minimizing percent differences in the weights rather than absolute differences?

The above gives you an idea of some of the issues that will be on our "What next" list. So stay tuned!

Afterword

We would like to thank Chip Alexander for his insightful discussion comments on this paper. His own research, albeit in another setting, certainly parallels ours. We are also grateful for the two references he mentioned that we had not seen: To his own 1990 work appearing in the ARC Proceedings and to the paper by Jayasuriya and Valliant, given in Orlando Thursday, after our paper was delivered. Both will be of help in handling our list of "What Nexts."

References

Burton, R. (1989), Unpublished Memorandum, National Center for Education Statistics.

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

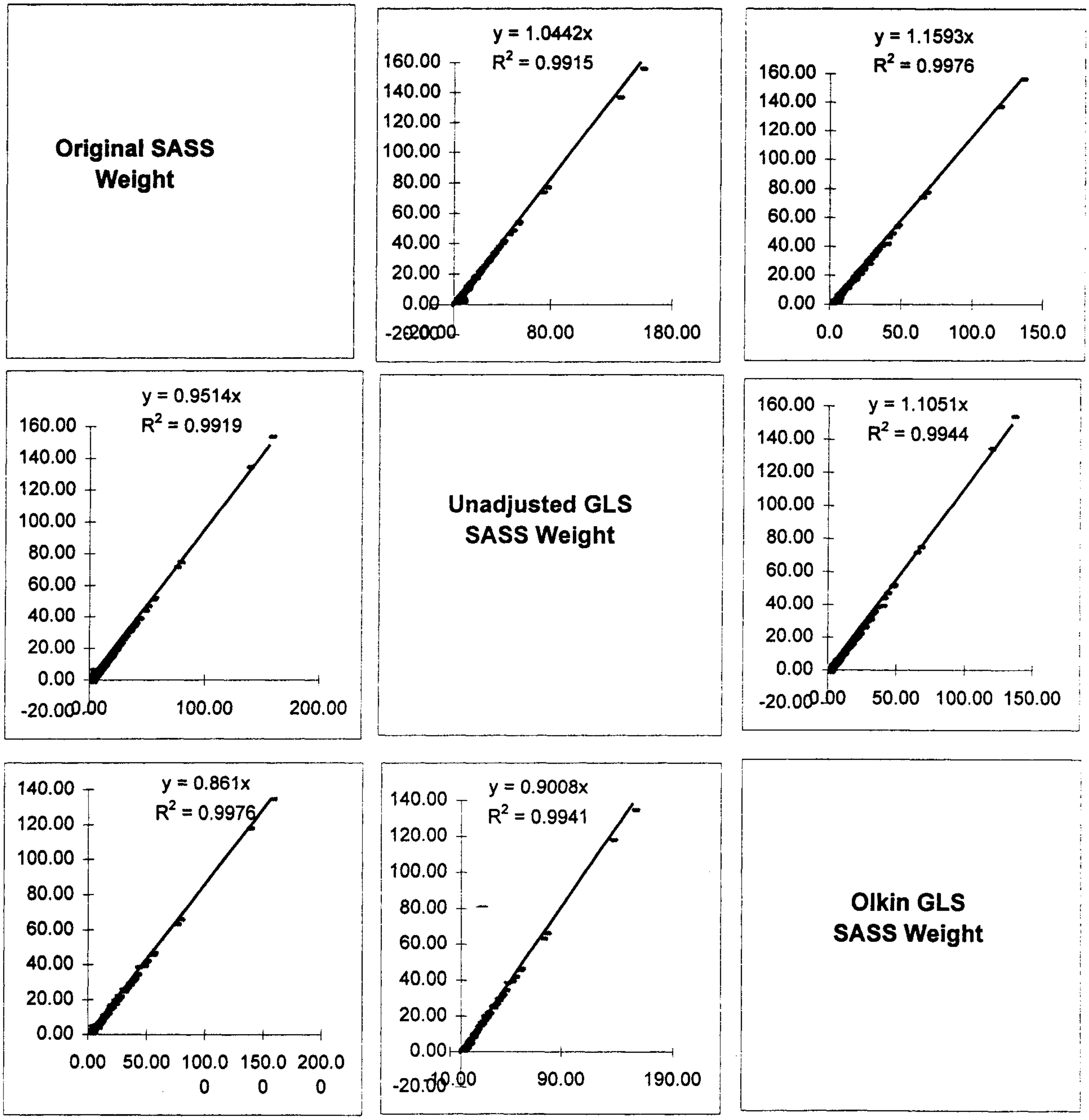
Deville, J.C., Särndal, C.E. and Sautory,O. (1993), "Generalized Raking Procedures in Survey Sampling," *Journal of the American Statistical Association,* 88, 1013-1020.

Holt, A., Kaufman, S., Scheuren, F. and Smith, W. (1994), "Intersurvey Consistency in School Surveys," Paper presented at the 1994 Joint Statistical Meetings, Toronto, Canada.

Imbens, G.W. and Hellerstein, J.K. (1993), "Raking and Regression," *Discussion Paper Number 1658*, Cambridge, MA, Harvard Institute of Economic Research, Harvard University.

Olkin, I. (1958), "Multivariate Ratio Estimation for Finite Populations," *Biometrika*, 45, 154-165.

Figure 2 -- Nonsectarian Regular
School weights, Unadjusted GLS, and Olkin-GLS SASS Compared



SOURCE: U.S.Department of Education, NCES, Private School of Schools and Staffing Surveys:1993-94, Private School Surveys, 1993-94