Calibration Estimators Including Fractional Exponents of Auxiliary Variables

Takis Merkouris*

Abstract

Auxiliary variables are extensively used in survey sampling to construct Generalized Regression (GR) estimators, or Optimal Regression (OR) estimators, of totals and means. These estimators are also calibration estimators, reproducing for the auxiliary variables the estimated parameters when the latter are available from external sources. This paper explores the possibility of improving the efficiency of such estimators when the utilized auxiliary variables are continuous, by augmenting the set of these variables with selected exponents of them. It is shown for the case of a single continuous auxiliary variable and simple random sampling or stratified simple random sampling, that the addition of a fractional exponent of the auxiliary variable improves the efficiency of the OR estimator to the degree of the implied increase of the coefficient of determination for the study variables. A simulation study shows that this additional variable improves the efficiency of the GR estimator greatly, even when the coefficient of determination is not increased.

Key Words: Continuous auxiliary variable, calibration, generalized regression estimator, optimal regression estimator.

1. Introduction

In survey sampling, auxiliary variables are extensively used at the estimation stage to improve the efficiency of estimators of interest, primarily of population totals and means. This is done ordinarily through generalized regression (GR), or through optimal regression (OR) for sampling designs for which this is possible.

As well known (see, for example, Deville, J. C., and Särndal (1992), Rao (1994), Andersson and Thorburn (2005)), both GR and OR are calibration procedures, whereby the sampling weights are adjusted (calibrated) so that the resulting estimates for the totals of the auxiliary variables are equal to the corresponding population totals (these being available from external sources). For any other variable of interest, the calibrated weights can be used to derive an estimate of the total in the form of weighted sum of the variable values, analogous to the linear form of the basic Horvitz-Thompson (HT) estimate involving the sampling weights. In this paper we will use the calibration formulation of both GR and OR procedures.

This paper addresses the question whether the calibration procedure uses to the maximum the information provided by a continuous auxiliary variable, or more information can be extracted when exponents of this variable are included in the procedure. Adding the exponent of the auxiliary variable to the calibration constraint entails calibrating its estimated total to the corresponding population total, if this is readily available. As we should strive for a most efficient sampling design, we should also search for maximum estimation efficiency. This possibility of most efficient use of auxiliary information is explored here for a single continuous auxiliary variable, and within the framework of optimal regression estimation, where exact theoretical results with respect to efficiency gains can be derived.

It is shown for the case of a single continuous auxiliary variable and simple random sampling or stratified simple random sampling, that the addition of a fractional exponent of the auxiliary variable improves the efficiency of the OR estimator to the degree of the

*Athens University of Business and Economics, 76 Patision Street, Athens 10434, Greece

implied increase of the coefficient of determination for the study variables. A simulation study shows that this additional variable also improves the efficiency of the GR estimator greatly, even when the coefficient of determination is not increased.

2. Theoretical development

Let $U = \{1, \cdots, k, \cdots, N\}$ denote a finite population of N units, and let s denote a sample of size n drawn from the population U, using a sampling design that defines inclusion probability $\pi_k = P(k \in s)$ for unit $k \in U$, and joint inclusion probability $\pi_{kl} = P(k, l \in s)$ for units $k, l \in U$. Assuming that $\pi_k > 0$ for all $k \in U$, the design weight of unit $k \in s$ is $w_k = 1/\pi_k$. For any variable of interest y, with values y_k, $k \in U$, and population total $Y = \sum_U y_k$, the Horvitz-Thompson (HT) estimator of Y is defined as $\hat{Y} = \sum_s w_k y_k$.

Let now $\mathbf{x}$ denote a $p \times 1$ vector of auxiliary variables, with known vector of population totals $\mathbf{X}$. The generalized regression estimator (GR) of Y is defined as

$$\hat{Y}^{GR} = \hat{Y} + \hat{\boldsymbol{\beta}}(\mathbf{X} - \hat{\mathbf{X}}) = \hat{Y} + \hat{\beta}_1(X_1 - \hat{X_1}) + \ldots + \hat{\beta}_p(X_p - \hat{X_p}),$$

where $\hat{\mathbf{X}} = \sum_s w_k \mathbf{x}_k$, and $\hat{\boldsymbol{\beta}} = \sum_s w_k y_k \mathbf{x}_k' (\sum_s w_k \mathbf{x}_k \mathbf{x}_k')^{-1}$. The optimal regression estimator (OR) of Y is defined as

$$\hat{Y}^{OR} = \hat{Y} + \hat{\boldsymbol{\beta}}^o(\mathbf{X} - \hat{\mathbf{X}}),$$

where $\hat{\boldsymbol{\beta}}^o = Cov(\hat{Y}, \hat{\mathbf{X}})(Var(\hat{\mathbf{X}}))^{-1}$ is the optimal (variance minimizing) regression coefficient. Both $\hat{Y}^{GR}$ and $\hat{Y}^{OR}$ are calibration estimators, that is, $\hat{\mathbf{X}}^{GR} = \hat{\mathbf{X}}^{OR} = \mathbf{X}$.

For a univariate auxiliary variable x, we consider its exponentiation $z = x^m$, where m may be any positive number, and assume that the corresponding population total $Z = \sum_U x_k^m$ is known. Then, the GR and OR estimators involving both of these two variables are given, respectively, by

$$\hat{Y}^{GR} = \hat{Y} + \hat{\beta}_1(X - \hat{X}) + \hat{\beta}_2(Z - \hat{Z}),$$

and

$$\hat{Y}^{OR} = \hat{Y} + \hat{\beta}_1^o(X - \hat{X}) + \hat{\beta}_2^o(Z - \hat{Z}). \tag{1}$$

The effect of including the variable z into the regression (calibration) estimation is possible to determine in the case of the OR estimator (1), which can be written (see Merkouris (2015)) as

$$\hat{Y}^{OR} = \hat{Y}^{OR|x} + \hat{\beta}_2^o(Z - \hat{Z}^{OR|x}), \tag{2}$$

where

$$\hat{Y}^{OR|x} = \hat{Y} + \frac{Cov(\hat{Y}, \hat{X})}{Var(\hat{X})}(X - \hat{X}), \qquad \hat{Z}^{OR|x} = \hat{Z} + \frac{Cov(\hat{Z}, \hat{X})}{Var(\hat{X})}(X - \hat{X})$$

are, respectively, the OR estimators of Y and Z using the single variable x, and the optimal partial regression coefficient $\hat{\beta}_2^o$ is given by

$$\hat{\beta}_2^o = \frac{Cov(\hat{Y}^{OR|x}, \hat{Z}^{OR|x})}{Var(\hat{Z}^{OR|x})}.$$

It easily follows that

$$Var(\hat{Y}^{OR|x}) = Var(\hat{Y}) - \frac{Cov^2(\hat{Y}, \hat{X})}{Var(\hat{X})}, \qquad Var(\hat{Z}^{OR|x}) = Var(\hat{Z}) - \frac{Cov^2(\hat{X}, \hat{Z})}{Var(\hat{X})}, \tag{3}$$

and

$$Cov(\hat{Y}^{OR|x}, \hat{Z}^{OR|x}) = Cov(\hat{Y}, \hat{Z}) - \frac{Cov(\hat{Y}, \hat{X})Cov(\hat{X}, \hat{Z})}{Var(\hat{X})}.$$

From (3), we can assess the variance reduction achieved by the OR estimation involving the variable x. Then the efficiency of the estimator $\hat{Y}^{OR|x}$ relative to $\hat{Y}$, measured by the relative difference eff$(\hat{Y}^{OR|x}, \hat{Y}) = [(Var(\hat{Y}) - Var(\hat{Y}^{OR|x})]/Var(\hat{Y})$, is readily verified to be eff$(\hat{Y}^{OR|x}, \hat{Y}) = \rho^2(\hat{Y}, \hat{X})$, that is, the square correlation coefficient of $\hat{Y}$ and $\hat{X}$.

It is now easy to show that

$$Var(\hat{Y}^{OR}) = Var(\hat{Y}^{OR|x}) - \frac{Cov^2(\hat{Y}^{OR|x}, \hat{Z}^{OR|x})}{Var(\hat{Z}^{OR|x})},$$

and hence determine the reduction of variability in the estimation of Y due to inclusion of the variable z in the optimal regression procedure. Furthermore, with straightforward algebra we obtain the efficiency of $\hat{Y}^{OR}$ relative to $\hat{Y}^{OR|x}$

$$\begin{aligned}
\text{eff}(\hat{Y}^{OR}, \hat{Y}^{OR|x}) &= \frac{Var(\hat{Y}^{OR|x}) - Var(\hat{Y}^{OR})}{Var(\hat{Y}^{OR|x})} \qquad (4)\\
&= \frac{Cov^2(\hat{Y}^{OR|x}, \hat{Z}^{OR|x})}{Var(\hat{Y}^{OR|x})Var(\hat{Z}^{OR|x})}\\
&= \frac{\left[Cov(\hat{Y}, \hat{Z})Var(\hat{X}) - Cov(\hat{Y}, \hat{X})Cov(\hat{X}, \hat{Z})\right]^2}{Var^2(\hat{X})Var(\hat{Y}^{OR|x})Var(\hat{Z}^{OR|x})}\\
&= \frac{\left[\rho(\hat{Y}, \hat{Z}) - \rho(\hat{Y}, \hat{X})\rho(\hat{X}, \hat{Z})\right]^2}{[1 - \rho^2(\hat{Y}, \hat{X})][1 - \rho^2(\hat{X}, \hat{Z})]}, \qquad (5)
\end{aligned}$$

where $\rho(\hat{Y}, \hat{Z})$ and $\rho(\hat{X}, \hat{Z})$ have similar to $\rho(\hat{X}, \hat{Y})$ meaning. It is clear from (5) that $\hat{Y}^{OR}$ is more efficient than $\hat{Y}^{OR|x}$ only if the three correlation coefficients satisfy the condition $\rho(\hat{Y}, \hat{Z}) - \rho(\hat{Y}, \hat{X})\rho(\hat{X}, \hat{Z}) \neq 0$. Noticing that equality holds if $\rho(\hat{X}, \hat{Z}) = 1$, a gain in efficiency is possible if $\rho(\hat{X}, \hat{Z})$ deviates from 1.

It follows that the efficiency of $\hat{Y}^{OR}$ relative to $\hat{Y}$, reflecting the compound optimal regression effect of the variables x and z, is

$$\begin{aligned}
\text{eff}(\hat{Y}^{OR}, \hat{Y}) &= \frac{Var(\hat{Y}) - Var(\hat{Y}^{OR})}{Var(\hat{Y})}\\
&= \frac{\rho^2(\hat{Y}, \hat{X}) + \rho^2(\hat{Y}, \hat{Z}) - 2\rho(\hat{Y}, \hat{X})\rho(\hat{Y}, \hat{Z})\rho(\hat{X}, \hat{Z})}{1 - \rho^2(\hat{X}, \hat{Z})}.
\end{aligned}$$

Under simple random sampling (SRS), $\rho(\hat{Y}, \hat{X})$, $\rho(\hat{Y}, \hat{Z})$ and $\rho(\hat{X}, \hat{Z})$ are equal, respectively, to the population correlation coefficients $\rho(y, x)$, $\rho(y, z)$ and $\rho(x, z)$. Then the efficiency eff$(\hat{Y}^{OR}, \hat{Y}^{OR|x})$ reduces to

$$\text{eff}(\hat{Y}^{OR}, \hat{Y}^{OR|x}) = \frac{[\rho(y, z) - \rho(y, x)\rho(x, z)]^2}{[1 - \rho^2(y, x)][1 - \rho^2(x, z)]},$$

which is the square partial correlation coefficient $\rho^2(y, z|x)$ between y and z controlling for x. Also, the efficiency eff$(\hat{Y}^{OR}, \hat{Y})$ reduces to

$$\text{eff}(\hat{Y}^{OR}, \hat{Y}) = \frac{\rho^2(y, x) + \rho^2(y, z) - 2\rho(y, x)\rho(y, z)\rho(x, z)}{1 - \rho^2(x, z)},$$

which is the coefficient of determination $R^2(y|x, z)$ in the regression of y on x and z.

3. Simulation Study

We have conducted a simulation study to assess the effect of adding $z = x^m$ in the calibration procedure that produces the OR and GR estimators. We simulated populations of size $N = 1000000$, consisting of the values of two variables y and x having the bivariate lognormal distribution with means $E(y) = 8$, $E(x) = 5$ and pairs of variances $Var(y) = (10, 50)$, $Var(x) = (10, 50)$, and associated coefficients of variation $CV(y) = (0.40, 0.88)$, $CV(x) = (0.63, 1.41)$. For each of these four bivariate distributions we specified three square correlations for (y, x), namely $\rho^2(y, x) = (0.25, 0.50, 0.75)$, thus creating 12 different populations of values of (y, x). Obviously, for each population the specification of $\rho^2(y, x)$ implies in turn the values of $\rho^2(y, z/x)$ and $R^2(y/x, z)$. Then, from each of these 12 populations we drew 20000 simple random samples of sizes $n = (3000, 1000, 300)$, and for each of the 36 simulated sampling settings we generated the HT estimate $\hat{Y}$ of the total Y, the OR estimates $\hat{Y}^{OR|x}$ and $\hat{Y}^{OR}$ (as defined in Section 2), and the analogously defined GR estimates $\hat{Y}^{GR|x}$ and $\hat{Y}^{GR}$. For $z = x^m$ we chose the fractional exponent $m = 1/4$ because it resulted in the highest efficiency gain eff$(\hat{Y}^{OR}, \hat{Y}^{OR|x})$ (highest $\rho^2(y, z/x)$) under the specified distribution of x, but also to avoid estimator instability associated with integer moments of x.

For each of the 36 sampling settings, using the 20000 samples we calculated the empirical relative (to the known Y) bias of all the above five estimators, the empirical counterparts of the OR efficiencies eff$(\hat{Y}^{OR|x}, \hat{Y})$, eff$(\hat{Y}^{OR}, \hat{Y})$, eff$(\hat{Y}^{OR}, \hat{Y}^{OR|x})$, and the same empirical efficiencies involving the GR estimators. As shown in Section 2, in SRS the OR efficiencies are equal to the correlation coefficients $\rho^2(y, x)$, $R^2(y/x, z)$ and $\rho^2(y, z/x)$, respectively, and thus their nominal values are set by the specified values of these coefficients for each of the four different bivariate distributions of (y, x) described above. These values are shown in Table 1, headed by the values $\rho^2(y, x) = (0.25, 0.50, 0.75)$ appearing in bold phase. Table 1 shows the empirical efficiencies of the OR and GR estimators for the various distribution and sample size settings.

We see in Table 1 that the inclusion of $z = x^m$ in the OR calibration procedure results in a nominal efficiency gain eff$(\hat{Y}^{OR}, \hat{Y}^{OR|x})$ (or $\rho^2(y, z/x)$) which increases as the correlation $\rho^2(y, x)$ increases, and as we move from distribution 1 to distribution 4 (to smaller $CV(y)$ relative to $CV(x)$), being negligible for distribution 1 and reaching the maximum value of 89.2% for correlation $\rho^2(y, x) = 0.75$ and distribution 4. This indicates that depending on the bivariate distribution of (y, x), strong correlation $\rho^2(y, x)$ may give substantial efficiency gains eff$(\hat{Y}^{OR}, \hat{Y}^{OR|x})$.

The empirical efficiencies of the estimators $\hat{Y}^{OR|x}$ and $\hat{Y}^{OR}$ are virtually identical to the nominal ones across the four distributions and the three correlation levels for sample size $n = 3000$, showing very small underestimation for size $n = 300$. The study showed (but it is not reported in Table 1) that whereas the estimator $\hat{Y}^{OR|x}$ had a very small bias for sizes $n = (1000, 300)$ in distributions 3 and 4 (with negligible effect on the mean square error), the estimator $\hat{Y}^{OR}$ had virtually zero bias.

The efficiency of the GR estimators $\hat{Y}^{GR|x}$ and $\hat{Y}^{GR}$, for of which there is no analytic expression, is empirically assessed relative to the efficiency of $\hat{Y}$ and relative to their OR counterparts.

The GR estimator $\hat{Y}^{GR|x}$ is very inefficient (relative to $\hat{Y}$) for distributions 2 and 4 and $\rho^2(y, x) = 0.25$ (by 35.3% and 82.6%, respectively), and although it improves substantially as $\rho^2(y, x)$ increases, it remains inefficient for all three correlation levels in distribution 4. In all other settings, $\hat{Y}^{GR|x}$ is more efficient than $\hat{Y}$, with the efficiency increasing as $\rho^2(y, x)$ increases. The estimator $\hat{Y}^{GR|x}$ is very inefficient relative to the estimator $\hat{Y}^{OR|x}$ in all settings, except for distribution 1 where its inefficiency is marginal. In all settings

this inefficiency lessens as $\rho^2(y,x)$ increases. Furthermore, the efficiency of $\hat{Y}^{GR|x}$ drops a little more than the efficiency of $\hat{Y}^{OR|x}$ as the sample size decreases. In the same settings where $\hat{Y}^{OR|x}$ showed a little bias, $\hat{Y}^{GR|x}$ showed a little more bias.

Regarding the effect of adding $z = x^{1/4}$ in the GR calibration procedure, Table 1 displays the most important finding of this empirical study: the estimator $\hat{Y}^{GR}$ is nearly as efficient as the estimator $\hat{Y}^{OR}$, irrespective of the distribution, the correlation $\rho^2(y,x)$ or the sample size. Moreover, the estimator $\hat{Y}^{GR}$ (like the estimator $\hat{Y}^{OR}$) has virtually zero bias.

We repeated the simulation with the same specifications as above, but now with the simulated populations stratified by the size of y, with five strata of sizes $N_1 = 400000$, $N_2 = 240000$, $N_3 = 200000$, $N_4 = 100000$ and $N_5 = 60000$. Simple random sampling in each stratum was used with uniform sample allocation $n_i = n/5$, $i = 1,\ldots,5$, giving a highly efficient design with inclusion probabilities $\pi_i = n_i/N_i = n/(5N)$, i.e., approximately proportional to the size of y. For this stratified random sampling (STRSRS) the nominal OR efficiencies eff$(\hat{Y}^{OR|x},\hat{Y})$, eff$(\hat{Y}^{OR},\hat{Y})$, eff$(\hat{Y}^{OR},\hat{Y}^{OR|x})$ are not equal to the specified correlation coefficients $\rho^2(y,x)$, $R^2(y/x,z)$ and $\rho^2(y,z/x)$, respectively, as they are in the case of SRS. They are significantly smaller. In Table 2, the values of the nominal OR efficiencies are given just below the values of $\rho^2(y,x)$, $R^2(y/x,z)$ and $\rho^2(y,z/x)$.

Table 2 shows that the empirical efficiencies of the estimators $\hat{Y}^{OR|x}$ and $\hat{Y}^{OR}$ are virtually identical to the nominal ones across the four distributions and the three correlation levels, although they drop very little or not at all as the sample size decreases.

The very small nominal OR efficiencies resulted in extremely inefficient GR estimator $\hat{Y}^{GR|x}$; this estimator is less efficient than $\hat{Y}$ in all settings. Another reason for the extreme inefficiency of $\hat{Y}^{GR|x}$ is that now the HT estimator $\hat{Y}$ is much more efficient because of the very efficient sampling design. Apparently, the regression effect is not added to the design effect. The inefficiency of $\hat{Y}^{GR|x}$ lessens only marginally as the correlation $\rho^2(y,x)$ increases. The estimator $\hat{Y}^{GR|x}$ is even more inefficient relative to the estimator $\hat{Y}^{OR|x}$, the latter being always more efficient than $\hat{Y}$.

With the addition of $z = x^{1/4}$ in the GR calibration procedure, the GR estimator $\hat{Y}^{GR}$ improves drastically, in all but one case, but is only better than the HT estimator $\hat{Y}$ in only one case. This is in sharp contrast with the unstratified case. One reason for this is that the nominal efficiency gain eff$(\hat{Y}^{OR},\hat{Y}^{OR|x})$ is in most settings small. In the only case that eff$(\hat{Y}^{OR},\hat{Y}^{OR|x})$ is large (75.5% for distribution 4), $\hat{Y}^{GR}$ shows impressive efficiency, approaching that of $\hat{Y}^{OR}$.

4. Conclusions

We have determined analytically the amount of the efficiency gain resulting from adding a fractional exponent $z = x^m$ of a continuous auxiliary variable x in the calibration procedure that generates the OR estimator of the total for a study variable y. Through an empirical study we showed that depending on the bivariate distribution of (y,x) and the sampling design, the efficiency gain may be substantial if the correlation of y with x is strong.

While the OR estimation is not feasible for many complex sampling designs, the GR estimation is always practicable (but not an analytic expression of its efficiency). The empirical study showed that adding a fractional exponent $z = x^m$ into the GR calibration procedure improves the efficiency of the GR estimator greatly, even when the efficiency gain for the OR estimator is only marginal.

Further research is needed on a theoretical explanation of the effect of $z = x^m$ on the

Table 1: Efficiency of OR and GR estimators (SRS)

	Population 1: CV(y)=0.88, CV(x)=0.63				Population 2: CV(y)=0.40, CV(x)=0.63				Population 3: CV(y)=0.88, CV(x)=1.41				Population 4: CV(y)=0.40, CV(x)=1.41			
	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R^2}(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\mid\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R^2}(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\mid\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R^2}(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\mid\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R^2}(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\mid\mathbf{x})$
		0.2500	0.2530	0.0040		**0.2500**	0.2697	0.0262		**0.2500**	0.3068	0.0758		**0.2500**	0.3647	0.1528
	3000				3000				3000				3000			
OR		0.2513	0.2535	0.0029		0.2542	0.2740	0.0265		0.2533	0.3093	0.0750		0.2537	0.3684	0.1537
GR		0.2302	0.2528	0.0294		-0.3531	0.2723	0.4622		0.1167	0.2997	0.2072		-0.8264	0.3671	0.6535
	1000				1000				1000				1000			
OR		0.2497	0.2498	0.0001		0.2439	0.2634	0.0258		0.2388	0.2922	0.0701		0.2415	0.3552	0.1499
GR		0.2273	0.2513	0.0311		-0.3951	0.2622	0.4712		0.0835	0.2851	0.2198		-0.9375	0.3542	0.6667
	300				300				300				300			
OR		0.2416	0.2359	-0.0076		0.2382	0.2546	0.0216		0.2330	0.2809	0.0625		0.2324	0.3469	0.1492
GR		0.22088	0.2424	0.0276		-0.3904	0.2559	0.4649		0.0683	0.2766	0.2236		-1.0685	0.3481	0.6848
		0.5000	0.5000	0.0000		**0.5000**	0.5246	0.0492		**0.5000**	0.5630	0.1259		**0.5000**	0.6802	0.3604
	3000				3000				3000				3000			
OR		0.5013	0.5008	-0.001		0.5036	0.5282	0.0496		0.5030	0.5648	0.1244		0.5023	0.6827	0.3624
GR		0.5021	0.5011	-0.002		0.1033	0.5225	0.4674		0.4127	0.5469	0.2286		-0.4400	0.6775	0.7761
	1000				1000				1000				1000			
OR		0.4994	0.4962	-0.0065		0.4963	0.5207	0.0485		0.4939	0.5524	0.1154		0.4954	0.6755	0.3568
GR		0.5009	0.4991	-0.0036		0.0754	0.5156	0.4761		0.3932	0.5374	0.2375		-0.5302	0.6710	0.7849
	300				300				300				300			
OR		0.4924	0.4833	-0.0179		0.4899	0.5118	0.0429		0.4903	0.5415	0.1004		0.4878	0.6678	0.3515
GR		0.4980	0.4912	-0.0135		0.0759	0.5087	0.4683		0.3851	0.5304	0.2363		-0.6479	0.6641	0.7962
		0.7500	0.7541	0.0164		**0.7500**	0.7740	0.0958		**0.7500**	0.8061	0.2244		**0.7500**	0.9730	0.8921
	3000				3000				3000				3000			
OR		0.7505	0.7545	0.0158		0.7520	0.7759	0.0963		0.7516	0.8071	0.2233		0.7494	0.9733	0.8934
GR		0.7391	0.7521	0.0502		0.4807	0.7682	0.5536		0.6903	0.7874	0.3136		-0.0956	0.9543	0.9583
	1000				1000				1000				1000			
OR		0.7498	0.7515	0.0069		0.7487	0.7726	0.0950		0.7489	0.8018	0.2108		0.7498	0.9728	0.8915
GR		0.7394	0.7511	0.0449		0.4659	0.7653	0.5605		0.6817	0.7841	0.3215		-0.1533	0.9542	0.9603
	300				300				300				300			
OR		0.7459	0.7444	-0.0064		0.7482	0.7717	0.0935		0.7482	0.7957	0.1887		0.7483	0.9717	0.8878
GR		0.7389	0.7468	0.0300		0.4717	0.7648	0.5548		0.6779	0.7810	0.3199		-0.2609	0.9531	0.9628

Table 2: Efficiency of OR and GR estimators (STRSRS, unequal probabilities)

	Population 1: CV(y)=0.88, CV(x)=0.63				Population 2: CV(y)=0.40, CV(x)=0.63				Population 3: CV(y)=0.88, CV(x)=1.41				Population 4: CV(y)=0.40, CV(x)=1.41			
	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R}^2(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\vert\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R}^2(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\vert\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R}^2(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\vert\mathbf{x})$	n	$\rho^2(\mathbf{y},\mathbf{x})$	$\mathbf{R}^2(\mathbf{y}/\mathbf{x},\mathbf{z})$	$\rho^2(\mathbf{y},\mathbf{z}\vert\mathbf{x})$
		0.2500	0.2530	0.0040		**0.2500**	0.2697	0.0262		**0.2500**	0.3068	0.0758		**0.2500**	0.3647	0.1528
		0.0304	0.0336	0.0033		0.0419	0.0697	0.0291		0.0336	0.0436	0.0103		0.0323	0.0912	0.0612
	3000				3000				3000				3000			
OR		0.0293	0.0327	0.0035		0.0452	0.0731	0.0292		0.0335	0.0421	0.0089		0.0314	0.0840	0.0543
GR		-3.5675	-1.9974	0.3437		-10.1740	-1.3580	0.7889		-4.6135	-1.6619	0.5258		-9.9916	-1.9591	0.7308
	1000				1000				1000				1000			
OR		0.0265	0.02600	-0.0005		0.0399	0.0649	0.0261		0.0322	0.0379	0.0058		0.0348	0.0872	0.0542
GR		-3.5659	-2.022	0.3379		-10.4109	-1.4189	0.7880		-4.7271	-1.7052	0.5276		-10.2707	-1.9541	0.7378
	300				300				300				300			
OR		0.0244	0.0165	-0.0081		0.0364	0.0522	0.0163		0.02911	0.02601	-0.0032		0.0331	0.0766	0.0449
GR		-3.5717	-2.0671	0.3291		-10.1644	-1.3965	0.7853		-5.0225	-1.8465	0.5273		-10.7509	-1.9357	0.7502
		0.5000	0.5000	0.0000		**0.5000**	0.5246	0.0492		**0.5000**	0.5630	0.1259		**0.5000**	0.6802	0.3604
		0.0887	0.0909	0.0024		0.1032	0.1748	0.0799		0.1215	0.1317	0.0117		0.0945	0.2439	0.1650
	3000				3000				3000				3000			
OR		0.0908	0.0927	0.0022		0.1025	0.1719	0.0774		0.1172	0.1261	0.0101		0.0976	0.2433	0.1615
GR		-2.0635	-2.0325	0.0101		-6.1621	-1.3362	0.6738		-2.8511	-1.2402	0.4183		-5.0295	-0.7967	0.7020
	1000				1000				1000				1000			
OR		0.0907	0.0915	0.0009		0.1010	0.1654	0.0716		0.1305	0.1342	0.0043		0.0972	0.2368	0.1546
GR		-2.0596	-2.0326	0.0088		-6.2058	-1.3588	0.6726		-2.8673	-1.2523	0.4176		-5.3290	-0.8598	0.7061
	300				300				300				300			
OR		0.0838	0.0749	-0.0097		0.1024	0.1619	0.0663		0.1232	0.1191	-0.0047		0.0909	0.2314	0.1545
GR		-2.0989	-2.0956	0.0011		-6.0145	-1.3183	0.6695		-3.0051	-1.3301	0.4182		-5.4845	-0.8569	0.7136
		0.7500	0.7541	0.0164		**0.7500**	0.7740	0.0958		**0.7500**	0.8061	0.2244		**0.7500**	0.9730	0.8921
		0.2368	0.2369	0.0001		0.2455	0.3675	0.1616		0.3499	0.3696	0.0302		0.3187	0.8331	0.7550
	3000				3000				3000				3000			
OR		0.2344	0.23303	-0.0018		0.2392	0.3654	0.1659		0.3412	0.3619	0.0315		0.3209	0.8357	0.7582
GR		-0.5671	-0.9402	-0.23807		-2.3432	-0.4275	0.5730		-0.8914	-0.1648	0.3842		-1.0375	0.7966	0.9002
	1000				1000				1000				1000			
OR		0.2279	0.2255	-0.0029		0.2403	0.3566	0.1531		0.3461	0.3595	0.0205		0.3193	0.8335	0.7555
GR		-0.5732	-0.9428	-0.2349		-2.4089	-0.4518	0.5741		-0.9195	-0.1893	0.3804		-1.0814	0.7918	0.8999
	300				300				300				300			
OR		0.2362	0.2272	-0.0118		0.2454	0.3594	0.1511		0.3529	0.3573	0.0067		0.3162	0.8299	0.7512
GR		-0.5612	-0.9368	-0.2406		-2.4445	-0.4589	0.5764		-0.9461	-0.2114	0.3775		-0.9991	0.7920	0.8959

efficiency of the GR estimator. Also, more detailed simulations involving other bivariate distributions of (y, x) and other sampling designs will help understanding this effect. In any real survey situation, the choice of the fractional exponent can be made empirically using the survey data. Moreover, using real survey data, an evaluation of the usefulness of such an extended GR calibration should also include the impact on estimates of many study variables, and on estimates for domains and for low-prevalence characteristics.

REFERENCES

Andersson, P. G., and Thorburn, D. (2005), "An Optimal Calibration Distance Leading to the Optimal Regression Estimator," *Survey Methodology*, 31, 95–99.

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

Merkouris, T. (2015), "An Efficient Estimation Method for Matrix Survey Sampling," *Survey Methodology*, 41, 237–262.

Rao, J. N. K. (1994), "Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage," *Journal of Official Statistics*, 10, 153–165.