Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Aligning Estimates for Common Variables in Two or More Sample Surveys

Robbert H. Renssen & Nico J. Nieuwenbroek

To cite this article: Robbert H. Renssen & Nico J. Nieuwenbroek (1997) Aligning Estimates for Common Variables in Two or More Sample Surveys, Journal of the American Statistical Association, 92:437, 368-374, DOI: 10.1080/01621459.1997.10473635

To link to this article: https://doi.org/10.1080/01621459.1997.10473635

Aligning Estimates for Common Variables in Two or More Sample Surveys

Robbert H. RENSSEN and Nico J. NIEUWENBROEK

In practice, for many sample surveys the estimates of several population totals are based on one set of weights, which reproduces the known population totals of auxiliary variables. Such a set can always be obtained by using the general regression estimator. If some variables, not necessarily with known population totals, are jointly collected in two sample surveys, then it may be desirable that the weights of both surveys produce the same estimates for the unknown population totals of the common variables. In this article we propose adjusting the general regression estimator to meet this consistency requirement by considering the common variables as additional auxiliary variables. It turns out that the adjusted general regression estimator generalizes Zieschang's method.

KEY WORDS: General regression estimators; Multivariate auxiliary information; Two-phase sampling; Weighting.

1. INTRODUCTION

If two sample surveys have some variables in common, then it may be attractive for both surveys to give virtually the same estimates for the population totals of these common variables. We use the term *common variables* for those variables observed in both surveys but for which the corresponding population totals are unknown. Typical examples of such variables include household characteristics such as size or composition of the household. For each survey, it is generally required that known population totals of auxiliary variables (i.e. control variables) be reproduced by weighting-type estimators. Often, these control variables are categorical and concern, for example sex, age, marital status, or region. For these variables, the known population totals are available as population counts. The requirement of reproducing these counts is always fulfilled if one takes these control variables as regressors in the general regression estimator (see, e.g., Särndal, Swensson, and Wretman 1992). For the common variables, we propose estimating the population totals by pooling both surveys, and then simultaneously using these common variables as additional regressors. The implicitly defined weights of this adjusted regression estimator reproduce the known population totals of the control variables, as well as the estimates of the population totals of the common variables. In this sense, we call the weights of the adjusted regression estimator reproductive with respect to the control variables and consistent with respect to the common variables.

A method for obtaining consistent weights with respect to the common variables of two sample surveys has already been proposed by Zieschang (1986, 1990). A generalization of a constrained minimum distance method is proposed to align estimates of comparable totals between the two surveys. The weighting method proposed herein is more general in that it reduces to that of Zieschang for a specific choice of the combined estimates of the population totals of the common variables. Alternative choices are available within the general framework developed. Furthermore, our method is written in a cleaner form, which enables us to address the choice of the combined estimates of the population totals of the common variables more clearly. Finally, our method easily can be extended to more than two surveys.

The adjusted general regression estimator can serve other important purposes beyond consistency. For instance, suppose that there exists evidence that a certain variable for which the population total is unknown is rather highly correlated with several important target variables. If that variable can be observed with minor costs, then it is worthwhile to include that variable in various surveys and use it as a common variable in the adjusted general regression estimator. In this application we are not interested in the population total of the common variable by itself, but we merely use the common variable as a tool to improve the estimates of the target variables. The elaboration of this proposal closely resembles two-phase sampling for the general regression estimator (see Särndal et al. 1992, chap. 9.7).

A promising application of the adjusted general regression estimator lies in its potential support of an attempt to decrease respondent burden and possibly increase the response rate by means of a split questionnaire survey design (see Raghunathan and Grizzle 1995). Sometimes a survey is originally projected with a questionnaire bearing many questions. It is expected that very long questionnaires will discourage potential respondents, resulting in high nonresponse rates. In such situations one may use shorter questionnaires, say form A and B, instead. Both forms contain questions about control variables and sensibly chosen questions to be used as the common variables. One part of the remaining questions is assigned to form A; the other part, to form B. By drawing two independent samples, form A can be assigned to the first sample and form B to the second. Now the theory of the adjusted general regression estimator can be applied to obtain consistent estimates with respect

Robbert H. Renssen is Senior Survey Methodologist and Nico J. Nieuwenbroek is Survey Methodologist, Department of Statistical Methods, Division of Research and Development, Statistics Netherlands, 6401 CZ Heerlen, The Netherlands. The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors wish to thank Hans Akkerboom, Ger Slootbeek, and Peter Kooiman for their careful reading and helpful comments. The authors also thank two anonymous referees and an associate editor for their valuable suggestions to improve this article.

to the common variables. Because the questionnaires are shorter, it is hoped that the nonresponse rates are smaller, resulting in less-biased estimates. Naturally, there is a loss of precision with respect to the target variables, due to the reduced number of observations in comparison to the original survey with the complete questionnaire. If the common variables are rather highly correlated with the target variables, then this loss of precision is limited.

Although, theoretically, we have found a solution for obtaining consistency among estimates between surveys, in practice complications may arise because common variables in the strict sense are not easily found, mainly due to discrepancies between definitions, methods of observation (e.g., primary versus secondary data sources), and reference periods. These complications can be reduced if the involved survey processes are harmonized at an early stage.

A disadvantage of the method is the increased possibility of negative weights, due to the enlarged number of explanatory variables. The occurrence of negative weights is inherent to the general regression estimator, and for many users this is an undesirable feature. The seriousness of negative weights depends on their number and magnitude. Several techniques are available to deal with negative weights for the general regression estimator (see, e.g., Deville and Särndal 1992 and Huang and Fuller 1978).

In Section 2 we briefly discuss the general regression estimator and introduce the notation. In Section 3 we define the adjusted general regression estimator. This is just the ordinary regression estimator with the common variables as additional regressors. By means of well-known theory about "introducing further regressors," this estimator is rewritten as the ordinary regression estimator plus an adjustment term. Inspired by the adjustment term, in Section 4 we suggest several alternative ways to estimate the population totals of the common variables. We elaborate on the adjusted general regression estimator and compare it to Zieschang's method. In section 5 we illustrate the efficiency gain obtainable with the adjusted general regression estimator using an example with data from an existing survey. In Section 6 we briefly discuss the extension of the adjusted general regression estimator to more than two surveys, the estimator's relation to two-phase sampling for the general regression estimator, and the use of calibration methods to deal with negative weights for the adjusted general regression estimator.

2. THE GENERAL REGRESSION ESTIMATOR

We consider a population U consisting of N elements $u_1, \ldots, u_N$, and associate with each element u_k a value y_k of a scalar target variable and a p vector $\mathbf{x}_k$ with values of p auxiliary variables. Let

$$\mathbf{Y}_{(N\times 1)} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

and

$$\mathbf{X}_{(N\times p)} = \begin{pmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_N^t \end{pmatrix},$$

where $\mathbf{x}_k^t$ denotes the transpose of $\mathbf{x}_k$. The population totals for all auxiliary variables are assumed to be known and are denoted by the p vector $\mathbf{t}_\mathbf{x}$; $\mathbf{t}_\mathbf{x} = \mathbf{X}^t\boldsymbol{\iota}$, where $\boldsymbol{\iota}$ is an N vector of 1s. We consider only sampling without replacement. For a given sample design, let π_k be the first-order inclusion probability of u_k and let π_{kl} be the second-order inclusion probability of u_k and u_l. Note that $\pi_{kl} = \pi_k$ for $k = l$. We assume that the first- and second-order inclusion probabilities are strictly positive. Let

$$\boldsymbol{\Pi}_{(N\times N)} = \operatorname{diag}(\pi_1, \ldots, \pi_N)$$

and

$$\boldsymbol{\Delta}_{(N\times N)} = (\delta_{kl}) \quad \text{with} \quad \delta_{kl} = \pi_{kl} - \pi_k\pi_l.$$

The elements of a sample S of size n can be determined by a matrix $\mathbf{T}$ of order $(n \times N)$, where

$$\mathbf{T}_{ij} = \begin{cases} 1 & \text{if in the } i\text{th draw element } u_j \text{ is selected,} \\ 0 & \text{otherwise.} \end{cases}$$

Now, the sample analogs of $\mathbf{Y}, \mathbf{X}, \boldsymbol{\Pi}$, and $\boldsymbol{\Delta}$ can be easily written as

$$\mathbf{Y}_S = \mathbf{T}\mathbf{Y}, \qquad \mathbf{X}_S = \mathbf{T}\mathbf{X}, \qquad \boldsymbol{\Pi}_S = \mathbf{T}\boldsymbol{\Pi}\mathbf{T}^t,$$

and

$$\boldsymbol{\Delta}_S = \mathbf{T}\boldsymbol{\Delta}\mathbf{T}^t.$$

We assume that $n > p$ and that $\mathbf{X}_S$ is of full rank p. Let $\boldsymbol{\iota}_S$ denote an n vector of 1s and let $\hat{y}_{\mathrm{HT}} = \mathbf{Y}_S^t\boldsymbol{\Pi}_S^{-1}\boldsymbol{\iota}_S$ and $\hat{\mathbf{x}}_{\mathrm{HT}} = \mathbf{X}_S^t\boldsymbol{\Pi}_S^{-1}\boldsymbol{\iota}_S$ be the Horvitz–Thompson estimators for $\mathbf{t}_y = \mathbf{Y}^t\boldsymbol{\iota}$ and $\mathbf{t}_\mathbf{x}$. The general regression estimator for $\mathbf{t}_y$ is defined as

$$\hat{y}_{\mathrm{R}} = \hat{y}_{\mathrm{HT}} + \hat{\mathbf{B}}^t(\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{HT}}), \tag{1}$$

where

$$\hat{\mathbf{B}} = (\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{X}_S)^{-1}\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{Y}_S$$

and $\boldsymbol{\Lambda}_S$ is some (symmetric) positive definite matrix of order $(n \times n)$. Särndal et al. (1992) suggested taking $\boldsymbol{\Lambda}_S = \boldsymbol{\Pi}_S^{-1}\boldsymbol{\Sigma}_S^{-1}$, where $\boldsymbol{\Sigma}_S = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^t$ and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_N^2)$. The σ_k^2 in $\boldsymbol{\Sigma}$ can be interpreted as the variance of independent random variables y_k defined in a superpopulation model, of which the y_k are supposed to be the outcomes. It is required that all σ_k^2 be known up to a common scale factor; that is, $\sigma_k^2 = v_k\sigma^2$ with $v_k (v_k > 0)$. Särndal et al. (1992) derived $\hat{B}$ as the sample estimator of the population parameter $\mathbf{B} = (\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y}$.

Let $\mathbf{I}$ denote the identity matrix. An important special case is $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$; that is, all variances $\sigma_k^2, k = 1, \ldots, N$ are equal. This results in the regression estimator of Bethlehem and Keller (1987). If the population elements represent households of size m_k and if we take $\sigma_k^2 = m_k\sigma^2$

(i.e., the variance of y_k is modeled to be proportional to the household size), then the general regression estimator corresponds to the GLS-P method of Alexander (1987). Lemaître and Dufour (1987) arrived at essentially the same estimator from a different point of view.

An alternative expression for the general regression estimator is given by

$$\begin{aligned}\hat{y}_{\mathrm{R}} &= \mathbf{Y}_S^t[\boldsymbol{\Pi}_S^{-1}\iota_S + \boldsymbol{\Lambda}_S\mathbf{X}_S(\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{X}_S)^{-1}(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{x}}_{\mathrm{HT}})] \\ &= \mathbf{Y}_S^t\mathbf{W}_S,\end{aligned}$$

from which it is obvious that the general regression estimator implicitly defines an n vector of weights $\mathbf{W}_S$. We note that $\mathbf{X}_S^t\mathbf{W}_S = \mathbf{t}_{\mathbf{x}}$. Thus, applying these weights to any of the auxiliary variables will yield the corresponding population total.

By a Taylor linearization, it can be shown (see, e.g., Särndal et al. 1992, result 6.6.1) that $\hat{y}_{\mathrm{R}}$ is approximately design unbiased with approximated design variance

$$\mathrm{AV}(\hat{y}_{\mathrm{R}}) = (\mathbf{Y} - \mathbf{XB})^t\boldsymbol{\Pi}^{-1}\boldsymbol{\Delta}\boldsymbol{\Pi}^{-1}(\mathbf{Y} - \mathbf{XB}). \quad (2)$$

This variance can be estimated by

$$v(\hat{y}_{\mathrm{R}}) = (\mathbf{Y}_S - \mathbf{X}_S\hat{\mathbf{B}})^t\boldsymbol{\Pi}_S^{-1}\check{\boldsymbol{\Delta}}_S\boldsymbol{\Pi}_S^{-1}(\mathbf{Y}_S - \mathbf{X}_S\hat{\mathbf{B}}), \quad (3)$$

where $\check{\boldsymbol{\Delta}}_S = \mathbf{T}\check{\boldsymbol{\Delta}}\mathbf{T}^t$ and $\check{\boldsymbol{\Delta}} = (\check{\delta}_{kl})$ with $\check{\delta}_{kl} = \delta_{kl}/\pi_{kl}$.

The general regression estimator has been extensively discussed by Särndal et al. (1992). They derived the general regression estimator using a model-assisted approach. Bethlehem and Keller (1987) derived it using a design-based approach. Luery (1986) showed that the general regression estimator can be obtained from a constrained minimum distance method, called generalized least squares (GLS) by Zieschang (1986) in an application with respect to the Consumer Expenditure Survey. The general regression estimator belongs to the more general class of QR estimators, first defined by Wright (1983). Wright (1983) gave sufficient conditions under which QR estimators are asymptotically design unbiased (ADU). Sufficient conditions for the consistency of the general regression estimator were given by Isaki and Fuller (1982) and Robinson and Särndal (1983). The latter authors also gave sufficient conditions for the general regression estimator to be ADU.

3. THE COMMON VARIABLES USED AS ADDITIONAL REGRESSORS

Suppose that two sample surveys have several variables in common. If the population totals of these variables are known, then the variables can be used as control variables in the general regression estimator for either survey. If the population totals are unknown, then we may estimate them and use these common variables with their estimated totals, jointly with the control variables, as regressors in the general regression estimators for both surveys. Then the implicitly defined weights of the resulting estimators will reproduce the estimated population totals of the common variables. In this section we consider the adjusted regression estimator for one sample, which we call S; that of the other sample, S', can be treated analogously.

We adjust the general regression estimator (1) by including q common variables with estimated population totals. Associate with each element u_k a q vector $\mathbf{z}_k$ with values of the q common variables. Let the q vector $\mathbf{t}_{\mathbf{z}}$ denote the true population totals of these variables; that is, $\mathbf{t}_{\mathbf{z}} = \mathbf{Z}^t\iota$, where

$$\mathbf{Z}_{(N\times q)} = \begin{pmatrix} \mathbf{z}_1^t \\ \vdots \\ \mathbf{z}_N^t \end{pmatrix}.$$

The Horvitz–Thompson estimator for $\mathbf{t}_{\mathbf{z}}$ is denoted by $\hat{\mathbf{z}}_{\mathrm{HT}} = \mathbf{Z}_S^t\boldsymbol{\Pi}_S^{-1}\iota_S$, where $\mathbf{Z}_S$ is the sample analog of $\mathbf{Z}$. The adjusted regression estimator is given by

$$\hat{y}_{\mathrm{AR}} = \hat{y}_{\mathrm{HT}} + \hat{\mathbf{B}}_{\mathrm{A}}^t(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{x}}_{\mathrm{HT}}) + \hat{\mathbf{D}}_{\mathrm{A}}^t(\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathrm{HT}}), \quad (4)$$

where $\hat{\mathbf{B}}_{\mathrm{A}}$ and $\hat{\mathbf{D}}_{\mathrm{A}}$ are simultaneously obtained from

$$\begin{pmatrix} \hat{\mathbf{B}}_{\mathrm{A}} \\ \hat{\mathbf{D}}_{\mathrm{A}} \end{pmatrix} = ((\mathbf{X}_S\mathbf{Z}_S)^t\boldsymbol{\Lambda}_S(\mathbf{X}_S\mathbf{Z}_S))^{-1}(\mathbf{X}_S\mathbf{Z}_S)^t\boldsymbol{\Lambda}_S\mathbf{Y}_S \quad (5)$$

and $\hat{\mathbf{z}}$ is a q vector of estimates of $\mathbf{t}_{\mathbf{z}}$. (We discuss some estimators for $\hat{\mathbf{z}}$ in the next section.) We assume that $n > p + q$ and that the partitioned matrix $(\mathbf{X}_S\mathbf{Z}_S)$ is of full rank $p+q$. We note that (4) can be interpreted as the general regression estimator, where the population totals of a number of auxiliary variables are unknown and replaced by certain estimates. We now state a fundamental property of the adjusted general regression estimator (4), which is proven in Appendix A. Let

$$\begin{aligned}\hat{y}_{\mathrm{R}} &= \hat{y}_{\mathrm{HT}} + \hat{\mathbf{B}}^t(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{x}}_{\mathrm{HT}}) \quad \text{with} \\ \hat{\mathbf{B}} &= (\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{X}_S)^{-1}\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{Y}_S\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{z}}_{\mathrm{R}} &= \hat{\mathbf{z}}_{\mathrm{HT}} + \hat{\mathbf{L}}^t(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{x}}_{\mathrm{HT}}) \quad \text{with} \\ \hat{\mathbf{L}} &= (\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{X}_S)^{-1}\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{Z}_S\end{aligned}$$

be the general regression estimators for the population totals $\mathbf{t}_y$ and $\mathbf{t}_{\mathbf{z}}$. We assume that $\mathbf{X}_S$ is of full rank $p \geq 1$. If $(\mathbf{X}_S\mathbf{Z}_S)$ is of full rank $p+q$, then the adjusted general regression estimator $\hat{y}_{\mathrm{AR}}$ given by (4) can be rewritten as

$$\hat{y}_{\mathrm{AR}} = \hat{y}_{\mathrm{R}} + \hat{\mathbf{D}}_{\mathrm{A}}^t(\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathrm{R}}). \quad (6)$$

The partial regression coefficient $\hat{\mathbf{D}}_{\mathrm{A}}$ can be written as

$$\hat{\mathbf{D}}_{\mathrm{A}} = (\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Y}_S$$

with

$$\mathbf{R}_S = \boldsymbol{\Lambda}_S - \boldsymbol{\Lambda}_S\mathbf{X}_S(\mathbf{X}_S^t\boldsymbol{\Lambda}_S\mathbf{X}_S)^{-1}\mathbf{X}_S^t\boldsymbol{\Lambda}_S.$$

We see that the adjusted general regression estimator $\hat{y}_{\mathrm{AR}}$ is equal to the general regression estimator $\hat{y}_{\mathrm{R}}$ plus an adjustment term. This adjustment term can be viewed as an attempt to further improve the general regression estimator. However, and probably more important, it is a means to achieve consistent results between the two samples with

respect to the common variables. The adjusted general regression estimator $\hat{y}_{AR}$ given by (6) implicitly defines the following n vector of weights:

$$\mathbf{W}_{AS} = \mathbf{W}_S + \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t R_S\mathbf{Z}_S)^{-1}(\hat{\mathbf{z}} - \hat{\mathbf{z}}_R), \quad (7)$$

where

$$\mathbf{W}_S = \mathbf{\Pi}_S^{-1}\iota_S + \mathbf{\Lambda}_S\mathbf{X}_S(\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S)^{-1}(\mathbf{t}_x - \hat{\mathbf{x}}_{HT})$$

is the n vector of weights corresponding to the general regression estimator $\hat{y}_R$. Noting that $\mathbf{X}_S^t\mathbf{R}_S = \mathbf{0}$, it is readily seen that indeed $\mathbf{X}_S^t\mathbf{W}_{AS} = \mathbf{X}_S^t\mathbf{W}_S = \mathbf{t}_x$ and $Z_S^t\mathbf{W}_{AS} = \hat{\mathbf{z}}$.

A special case to consider is when no control variables are used; that is, $p = 0$. Then the adjusted regression estimator is simply

$$\hat{y}_{AR} = \hat{y}_{HT} + \hat{\mathbf{D}}^t(\hat{\mathbf{z}} - \hat{\mathbf{z}}_{HT})$$

with

$$\hat{\mathbf{D}} = (\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S)^{-1}\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Y}_S.$$

We note that this expression resembles the expression given by (1), with the difference that population totals for the regressors must be estimated instead of being known.

The result given here is an instance of relationships that generally hold between partial regressors in the general linear model. There is an extensive body of literature about "introducing further regressors" in linear regression analysis (see Seber 1977).

4. PROPERTIES OF THE ADJUSTED GENERAL REGRESSION ESTIMATOR

An important issue is estimating the unknown population totals of the common variables. A natural choice is

$$\hat{\mathbf{z}} = \mathbf{P}\hat{\mathbf{z}}_R + \mathbf{Q}\hat{\mathbf{z}}'_R,$$

where $\mathbf{P}$ and $\mathbf{Q}$ are two matrices of order $(q \times q)$ such that $\mathbf{P} + \mathbf{Q} = \mathbf{I}$, and $\hat{\mathbf{z}}'_R$ is the general regression estimator for $\mathbf{t}_z$ from the second sample, analogously defined as $\hat{\mathbf{z}}_R$. Here we discuss some choices for $\mathbf{P}$ and $\mathbf{Q}$.

First, take either $\mathbf{P}$ or $\mathbf{Q}$ to consist of 0s, so that $\hat{\mathbf{z}} = \hat{\mathbf{z}}'_R$ or $\hat{\mathbf{z}} = \hat{\mathbf{z}}_R$. In this case one survey is declared the standard and the other is adjusted.

The choice $\mathbf{P} = (n_1 + n_2)^{-1}n_1\mathbf{I}$ and $\mathbf{Q} = (n_1 + n_2)^{-1} \times n_2\mathbf{I}$ takes into account the difference in sample size. This choice and the first one are both special cases of $\mathbf{P} = \gamma\mathbf{I}$ and $\mathbf{Q} = (1 - \gamma)\mathbf{I}$, where $\gamma, 0 \le \gamma \le 1$, is a crude measure of the amount of confidence in one estimator compared to the other. The choice of γ may depend on indicators for several survey errors, such as frame errors, sampling errors, nonresponse errors, and measurement errors. (We refer to Groves 1989 for an extensive discussion about survey errors.)

Let $V(\hat{\mathbf{z}}_R)$ be the covariance matrix of $\hat{\mathbf{z}}_R$ and $V(\hat{\mathbf{z}}'_R)$ that of $\hat{\mathbf{z}}'_R$. Then the choice

$$\mathbf{P} = V(\hat{\mathbf{z}}'_R)[V(\hat{\mathbf{z}}_R) + V(\hat{\mathbf{z}}'_R)]^{-1}$$

and

$$\mathbf{Q} = V(\hat{\mathbf{z}}_R)[V(\hat{\mathbf{z}}_R) + V(\hat{\mathbf{z}}'_R)]^{-1}$$

minimizes the variance of $\mathbf{a}^t\hat{\mathbf{z}}$ for an arbitrary q vector $\mathbf{a}$. In practice, these covariance matrices are unknown and must be replaced by their estimates, which can be obtained from (3). The resulting matrices are estimates for $\mathbf{P}$ and $\mathbf{Q}$ and hence are denoted by $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$.

Taking into account that the matrices $\mathbf{P}$ and $\mathbf{Q}$ may be unknown and must be estimated, we insert $\hat{\mathbf{z}} = \hat{\mathbf{P}}\hat{\mathbf{z}}_R + \hat{\mathbf{Q}}\hat{\mathbf{z}}'_R$ into (6). This gives

$$\hat{y}_{AR} = \hat{y}_R + \hat{\mathbf{D}}_A^t\hat{\mathbf{Q}}(\hat{\mathbf{z}}'_R - \hat{\mathbf{z}}_R). \quad (8)$$

The corresponding weight vector $\mathbf{W}_{AS}$ given by (7) becomes

$$\mathbf{W}_{AS} = \mathbf{W}_S + \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\hat{\mathbf{Q}}(\hat{\mathbf{z}}'_R - \hat{\mathbf{z}}_R). \quad (9)$$

The first term $\mathbf{W}_S$ balances the sample toward the known population totals $\mathbf{t}_x$. The second term provides the consistency requirement with the other sample.

The weights given by (9) are equal to those implicitly defined by the GLS method of Zieschang (1990) if $\hat{\mathbf{Q}}$ has the particular form

$$\hat{\mathbf{Q}} = (\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S + \mathbf{Z}_{S'}^t\mathbf{R}_{S'}\mathbf{Z}_{S'})^{-1}, \quad (10)$$

where the subscript S' denotes the other sample. This is demonstrated in Appendix B. It is implicitly assumed that at least one control variable is used in both surveys. The $(q \times q)$ matrix $\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S$ is equal to

$$\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S = (\mathbf{Z}_S - \mathbf{X}_S\hat{\mathbf{L}}_S)^t\mathbf{\Lambda}_S(\mathbf{Z}_S - \mathbf{X}_S\hat{\mathbf{L}}_S).$$

An analogous expression can be obtained for $\mathbf{Z}_{S'}^t\mathbf{R}_{S'}\mathbf{Z}_{S'}$. We note that for $q = 1$, these expressions are weighted sums of squared residuals. The weights are given by $\mathbf{\Lambda}_S$ for the first sample and by $\mathbf{\Lambda}_{S'}$ for the second sample. However, $\mathbf{\Lambda}_S$ also determines the regression coefficient of the first sample, as does $\mathbf{\Lambda}_{S'}$ for the second sample. This dependence may be too restrictive in many circumstances. For example, within the context of the model-assisted approach of Särndal et al. (1992), we have $\mathbf{\Lambda}_S = \mathbf{\Pi}_S^{-1}\mathbf{\Sigma}_S^{-1}$ and $\mathbf{\Lambda}_{S'} = \mathbf{\Pi}_{S'}^{-1}\mathbf{\Sigma}_{S'}^{-1}$. If for both samples the same control variables are used and the same $\mathbf{\Sigma}$ matrix are assumed, then

$$\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S = \sum_{k\in S}(\mathbf{z}_k - \mathbf{x}_k^t\hat{\mathbf{L}}_S)^2/\pi_k\sigma_k^2$$

and

$$\mathbf{Z}_{S'}^t\mathbf{R}_{S'}\mathbf{Z}_{S'} = \sum_{k\in S'}(\mathbf{z}_k - \mathbf{x}_k^t\hat{\mathbf{L}}_{S'})^2/\pi'_k\sigma_k^2.$$

Clearly, both weighted sums of squared residuals are estimators for the same population parameter. Inserting these estimators into (10) gives a scalar $\hat{\mathbf{Q}}$, which can be considered to be an estimator for 1/2. Obviously, such a choice for $\mathbf{\Lambda}_S$ and $\mathbf{\Lambda}_{S'}$ does not account for differences in sample size.

The derivation of the approximated variance of $\hat{y}_{AR}$ given by (7) is similar to that for the general regression estimator. Assuming independence between the samples, the Taylor linearization technique yields

$$\mathrm{AV}(\hat{y}_{AR}) = \mathrm{AV}(\hat{y}_R) + \mathbf{D}_A^t\mathbf{Q}[\mathrm{AV}(\hat{\mathbf{z}}_R) + \mathrm{AV}(\hat{\mathbf{z}}'_R)]\mathbf{Q}^t\mathbf{D}_A - 2\mathbf{D}_A^t\mathbf{Q}[\mathrm{AC}(\hat{\mathbf{z}}_R, \hat{y}_R)],$$

Table 1. Different Cases Concerning Sample Sizes and Use of Control Variables

	Sample 1	*Sample 2*
Case 1		
Sample size	1,523	5,721
Target variables	Value self-occupied houses	Household income
Common variables	Ownership (1 category) and household size (5 categories)	
Control variables	Sex by age (10 categories)	Sex by age (16 categories)
Case 2		
Sample size	3,622	3,622
Target variables	Value self-occupied houses	Household income
Common variables	Ownership (1 category) and household size (5 categories)	
Control variables	Sex by age (10 categories)	Sex by age (16 categories)
Case 3		
Sample size	1,523	5,721
Target variables	Value self-occupied houses	Household income
Common variables	Ownership (1 category) and household size (5 categories)	
Control variables	Sex by age (10 categories)	Sex by age (10 categories)

where the approximated variances and covariances of the general regression estimators can be obtained from (2) by inserting the appropriate sample estimates. Gains are made when the additional regressors are highly correlated with the target variables and when the estimates with respect to the additional regressors are highly accurate.

5. AN EXAMPLE

In this section we compare the approximated variances of the adjusted general regression estimator for different $\mathbf{P}$ and $\mathbf{Q}$ matrices. Three choices for these matrices are involved: namely proportional (i.e., both matrices are diagonal with entries proportional to the sample sizes), optimal (i.e., the choice that minimizes the variance of $\mathbf{a}^t\hat{\mathbf{z}}$ for an arbitrary q vector $\mathbf{a}$), and Zieschang's choice. We know that for all choices, the adjusted general regression estimator produces consistent weights with respect to the common variables, so from this viewpoint the choice is irrelevant. However, there are differences with respect to the approximated variances as well as with respect to the ease of computation. The proportional choice for $\mathbf{P}$ and $\mathbf{Q}$ depends only on the sample sizes. The optimal choice for $\mathbf{P}$ and $\mathbf{Q}$ is determined by the sample sizes, the use of control variables, and the efficiency of the design. Zieschang's choice is determined by the use of control variables only. By means of an example based on the data from a Dutch household survey (the Income Panel Survey 1989), we try to gain some insight into these differences. Among others, in this household survey the variables value of self-occupied houses, household income, household size, sex, and age are observed.

We distinguish three cases. In each case two simple random samples from the household survey are drawn. The cases differ only in terms of sample sizes and control variables, as shown in Table 1.

In the first sample we consider value of self-occupied houses as the target variable and in the second sample household income. Both samples have six dummy variables in common. One dummy variable concerns the ownership of the house, and five dummy variables concern the household size. In case 1 the sample sizes are $n_1 = 1{,}523$ and $n_2 = 5{,}721$ households. In the first sample we cross the 2 sex categories (male and female) by 5 age categories, giving 10 control variables; in the second sample we cross the 2 sex categories by 8 age categories, giving 16 control variables. In case 2 the sample sizes are $n_1 = n_2 = 3{,}622$ households. We cross the two sex categories by five age categories in the first sample and by eight age categories in the second sample. In case 3 the sample sizes are $n_1 = 1{,}523$ and $n_2 = 5{,}721$ households. In both samples we cross the two sex categories by five age categories.

We wish to compare the variances of the target variables value of self-occupied houses and household income. Therefore, we have calculated for each case and for both samples the variance of the Horvitz–Thompson estimator, the variance of the general regression estimator, and for different $\mathbf{P}$ and $\mathbf{Q}$ matrices the variances of the adjusted regression estimators. The person characteristics sex and age are transformed into household characteristics by the method of Lemaître and Dufour (1987). To facilitate the comparison between the different variances, the variances of the Horvitz–Thompson estimator are used as references and are set at 100. The results are summarized in Table 2. In the first sample the general regression estimator with 10 control variables is used in all cases. As expected, it appears that this estimator shows an improvement over the Horvitz–Thompson estimator. Similarly, the general regres-

Table 2. Estimated Variances of Various Estimators Relative to the Corresponding Estimated Variance of the Horvitz–Thompson Estimator

	Variances	*Variances*
	Value of self-occupied houses	Total income
Case 1		
General regression estimator	96	86
Adjusted regression estimator		
• Proportional to sample size	55	83
• Optimal	55	82
• Zieschang	59	92
Case 2		
General regression estimator	96	86
Adjusted regression estimator		
• Proportional to sample size	70	77
• Optimal	70	76
• Zieschang	70	76
Case 3		
General regression estimator	96	92
Adjusted regression estimator		
• Proportional to sample size	55	87
• Optimal	55	87
• Zieschang	60	99

sion estimator with the same 10 control variables performs better in the second sample, as we can see from case 3. In turn, the general regression estimator with 16 control variables performs better than the general regression estimator with 10 control variables, as we can see from cases 1 and 2.

The variance reduction from using the common variables as additional regressors largely depends on three factors: the size of the one sample compared to the size of the other sample, the choice of the $\mathbf{P}$ and $\mathbf{Q}$ matrix, and the partial correlation between the target variables on the one hand and the common variables on the other hand with the control variables held constant. In particular, we have compared the various $\mathbf{P}$ and $\mathbf{Q}$ matrices, taking into account differences in sample size and control variables. From Table 2, we see that in all cases and for both samples, the proportional values for $\mathbf{P}$ and $\mathbf{Q}$ give virtually the same results as the optimal values, despite the differences in the first two cases between the use of control variables. Obviously, these differences are not large enough to greatly affect the variances obtained by the proportional choice. In the first and third cases, the variances obtained by Zieschang's choice differ from those obtained by the optimal choice. Especially in these cases, the sample sizes differ substantially. Because Zieschang's choice does not take into account the different sample sizes, this choice results in higher variances. Originally, Zieschang's choice is based on a case study involving two surveys with roughly the same sample size. For that particular case, the choice can be motivated.

In practice it is quite common that the number of control variables used in the general regression estimator depend on the sample size; the larger the sample size, the larger the number of control variables we can use. If the sample sizes are nearly equal, then the numbers of control variables probably are nearly equal as well, and the three choices give virtually the same variances, as in case 2. If the sample size of the one sample is much larger than the size of the other sample, then the choice of $\mathbf{P}$ and $\mathbf{Q}$ should depend not only on the difference in sample size but, also on the difference in correlation between the control variables on the one hand and the common variables on the other hand. The optimal choice does this implicitly. The proportional choice probably results in higher variances, depending on the differences in correlation, and Zieschang's choice results in higher variances, due to the difference in sample size.

6. DISCUSSION

The adjusted general regression estimator given by (4) or (6) can be extended in a straightforward manner to establish consistency with respect to the common variables between more than two, say k, sample surveys. Take $\hat{\mathbf{z}} = \sum_{i=1}^{k} \hat{\mathbf{P}}_i \hat{\mathbf{z}}_{i\mathrm{R}}$ as an estimator for $\mathbf{t}_\mathbf{z}$, where $\hat{\mathbf{P}}_i$ are $(q \times q)$ matrices adding up to the identity matrix and $\hat{\mathbf{z}}_{i\mathrm{R}}$ is the general regression estimator of $\mathbf{t}_\mathbf{z}$ from the ith survey, and use this combined estimate for each survey. This guarantees consistency with respect to the common variables between these surveys. It is worthwhile to note that the amount of computation per survey remains limited. The extended weight vector $\mathbf{W}_{AS}$ can be calculated from the original weight vector $\mathbf{W}_S$ plus a term that just requires inversion of the $(q \times q)$ matrix $\mathbf{Z}_S^t \mathbf{R}_S \mathbf{Z}_S$ and calculation of $\hat{\mathbf{z}}$.

There is a strong relationship between the adjusted general regression estimator given by (4) or (6) and regression estimators for two-phase sampling (see Särndal et al. 1992, chap. 9.7). Let the vector $\mathbf{x}_k$ correspond to Särndal's $\mathbf{x}_{1k}$, which uses auxiliary values known for all units in the population, and let $\mathbf{z}_k$ correspond to Särndal's $\mathbf{x}_{2k}$, which are known for observations in the first phase sample only. Moreover, let $\hat{\mathbf{z}}$ denote the general regression estimator based on the first-phase sample, $\hat{y}_{\mathrm{HT}}$, let $\hat{\mathbf{x}}_{\mathrm{HT}}$ and $\hat{\mathbf{z}}_{\mathrm{HT}}$ denote the Horvitz–Thompson estimators based on the second-phase sample, and let $\hat{y}_{\mathrm{R}}$ and $\hat{\mathbf{z}}_{\mathrm{R}}$ denote the general regression estimators based on the second-phase sample. Then

$$\hat{y}_{\mathrm{HT}} + \hat{\mathbf{B}}_{\mathrm{A}}^t(\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{HT}}) + \hat{\mathbf{D}}_{\mathrm{A}}^t(\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathrm{HT}}) = \hat{y}_{\mathrm{R}} + \hat{\mathbf{D}}_{\mathrm{A}}^t(\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathrm{R}})$$

can be considered a general regression estimator for two-phase sampling. The partial regression coefficients $\hat{\mathbf{B}}_{\mathrm{A}}^t$ and $\hat{\mathbf{D}}_{\mathrm{A}}^t$, simultaneously obtained from (5), are based on the second-phase sample. We note that this expression is slightly different from equation (9.7.2) of Särndal et al. (1992). A practical application of the general regression estimator for two-phase sampling was given by Heerschop and Lieftinck-Koeijers (1991). They suggested improving the current estimator for the total number of unemployed persons at time t by using an auxiliary variable denoting whether a person is working or not at time $t-1$. The population total of this auxiliary variable is unknown and estimated by pooling two independent estimates, one based on the current sample with retrospective questions and one based on a previous sample with actual information at time $t-1$. In our conception this auxiliary variable is considered an additional regressor.

Deville and Särndal (1992) developed calibration estimation, which is especially interesting with respect to the problem of negative weights. A calibration estimator uses calibration weights w_k that are as close as possible, in terms of a distance measure, to the original sampling design weights $d_k = \pi_k^{-1}$ while also respecting a set of constraints $\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{x}$. If the distance measure is based on $(w_k - d_k)^2 \sigma_k^2 / d_k$, then the calibration estimator corresponds to the general regression estimator. Following Deville and Särndal (1992, case 7), this specific distance measure can be modified to bound the range of the calibration weights w_k (i.e., $Ld_k \leq w_k \leq Ud_k$ for some constants L and U) without much change in the original estimator. The bounded calibration can be applied to bound the weights $\mathbf{W}_{AS}$ given by (7) as follows. Apply the bounded calibration method to obtain bounded calibration weights instead of the weights given in $\mathbf{W}_S$. The input weights are the original sampling design weights, and the set of constraints concern the control variables only. Analogously, modify the weights given by $\mathbf{W}_{S'}$. Pool the resulting calibration estimates with respect to the common variables to obtain an estimator for $\mathbf{t}_\mathbf{z}$. Apply the bounded calibration method again to obtain bounded calibration weights instead of the weights

given by $\mathbf{W}_{AS}$. The input weights are the bounded calibration weights obtained by the first application, and the set of constraints concern both the control variables and the common variables. We note that the resulting calibration estimator equals the adjusted general regression estimator for $L = -\infty$ and $U = \infty$.

APPENDIX A: PROOF OF THE FUNDAMENTAL PROPERTY OF THE ADJUSTED GENERAL REGRESSION ESTIMATOR

The corresponding normal equations of the multiple regression coefficient $\begin{pmatrix}\hat{\mathbf{B}}_A\\ \hat{\mathbf{D}}_A\end{pmatrix}$ given by (5) are

$$(\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S)\hat{\mathbf{B}}_A + (\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S)\hat{\mathbf{D}}_A = \mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{Y}_S$$

and

$$(\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{X}_S)\hat{\mathbf{B}}_A + (\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S)\hat{\mathbf{D}}_A = \mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Y}_S.$$

The upper equation gives $\hat{\mathbf{B}}_A = \hat{\mathbf{B}} - \hat{\mathbf{L}}\hat{\mathbf{D}}_A$. Inserting this expression into the lower equation, we obtain

$$\hat{\mathbf{D}}_A = (\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Y}_S,$$

and inserting this expression into (4), we obtain (6). The inverse of $\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S$ exists because $\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S$ and $(\mathbf{X}_S\mathbf{Z}_S)^t\mathbf{\Lambda}_S(\mathbf{X}_S\mathbf{Z}_S)$ are nonsingular. Indeed, the nonsingularity of $\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S$ implies the validity of the following well-known identity about partitioned matrices:

$$\det[(\mathbf{X}_S\mathbf{Z}_S)^t\mathbf{\Lambda}_S(\mathbf{X}_S\mathbf{Z}_S)] = \det[\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S]\det[\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S]$$

(see, e.g., Rao 1973, chap. 1, prob. 2.4), which immediately yields the nonsingularity of $\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S$.

APPENDIX B: THE GENERALIZED LEAST SQUARES METHOD

Here we show that the weights given by (9) with $\hat{\mathbf{Q}}$ given by (10) are equal to those given by Zieschang (1990, form. 3.10). To that end, we express both weight vectors in a similar form. Now on the one hand, the weights given by (9) can be written as

$$\begin{aligned}\mathbf{W}_{AS} &= \mathbf{W}_S + \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\hat{\mathbf{Q}}(\hat{\mathbf{z}}'_R - \hat{\mathbf{z}}_R)\\ &= \mathbf{W}_S - \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\hat{\mathbf{Q}}\hat{\mathbf{L}}^t(\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{HT})\\ &\quad + \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\hat{\mathbf{Q}}\hat{\mathbf{L}}'^t(\mathbf{t}'_\mathbf{x} - \hat{\mathbf{x}}'_{HT})\\ &\quad + \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S)^{-1}\hat{\mathbf{Q}}(\hat{\mathbf{z}}'_{HT} - \hat{\mathbf{z}}_{HT}).\end{aligned}$$

On the other hand, the weights given by Zieschang (1990) can be written, in terms of our notation, as

$$\begin{aligned}\hat{\mathbf{W}}_1 &= \mathbf{\Pi}_S^{-1}\iota_S + (\mathbf{\Lambda}_S\mathbf{X}_S \quad 0 \quad \mathbf{\Lambda}_S\mathbf{Z}_S)\\ &\quad\times\begin{pmatrix}\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S & 0 & \mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S\\ 0 & \mathbf{X}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{X}_{S'} & -\mathbf{X}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{Z}_{S'}\\ \mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{X}_S & -\mathbf{Z}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{X}_{S'} & \mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S + \mathbf{Z}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{Z}_{S'}\end{pmatrix}^{-1}\\ &\quad\times\begin{pmatrix}\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{HT}\\ \mathbf{t}'_\mathbf{x} - \hat{\mathbf{x}}'_{HT}\\ \hat{\mathbf{z}}'_{HT} - \hat{\mathbf{z}}_{HT}\end{pmatrix}\\ &= \mathbf{\Pi}_S^{-1}\iota_S + \mathbf{A}_1(\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{HT}) + \mathbf{A}_2(\mathbf{t}'_\mathbf{x} - \hat{\mathbf{x}}'_{HT}) + \mathbf{A}_3(\hat{\mathbf{z}}'_{HT} - \hat{\mathbf{z}}_{HT}),\end{aligned}$$

where the submatrices $\mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{A}_3$ are implicitly defined. They can be found by solving the following system of normal equations:

$$\mathbf{A}_1(\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S) + \mathbf{A}_3(\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{X}_S) = \mathbf{\Lambda}_S\mathbf{X}_S,$$

$$\mathbf{A}_2(\mathbf{X}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{X}_{S'}) - \mathbf{A}_3(\mathbf{Z}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{X}_{S'}) = \mathbf{0},$$

and

$$\mathbf{A}_1(\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S) - \mathbf{A}_2(\mathbf{X}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{Z}_{S'}) + \mathbf{A}_3(\mathbf{Z}_S^t\mathbf{\Lambda}_S\mathbf{Z}_S + \mathbf{Z}_{S'}^t\mathbf{\Lambda}_{S'}\mathbf{Z}_{S'}) = \mathbf{\Lambda}_S\mathbf{Z}_S.$$

From the first equation, it follows that $\mathbf{A}_1 = \mathbf{\Lambda}_S\mathbf{X}_S(\mathbf{X}_S^t\mathbf{\Lambda}_S\mathbf{X}_S)^{-1} - \mathbf{A}_3\hat{\mathbf{L}}^t$; from the second equation, we have $\mathbf{A}_2 = \mathbf{A}_3\hat{\mathbf{L}}'^t$. Inserting these values into the third equation gives $\mathbf{A}_3 = \mathbf{R}_S\mathbf{Z}_S(\mathbf{Z}_S^t\mathbf{R}_S\mathbf{Z}_S + \mathbf{Z}_{S'}^t\mathbf{R}_{S'}\mathbf{Z}_{S'})^{-1}$. Collecting results, we obtain

$$\begin{aligned}\hat{\mathbf{W}}_1 = \mathbf{W}_S &- \mathbf{A}_3\hat{\mathbf{L}}^t(\mathbf{t}_\mathbf{x} - \hat{\mathbf{x}}_{HT})\\ &+ \mathbf{A}_3\hat{\mathbf{L}}'^t(\mathbf{t}'_\mathbf{x} - \hat{\mathbf{x}}'_{HT}) + \mathbf{A}_3(\hat{\mathbf{z}}'_{HT} - \hat{\mathbf{z}}_{HT}).\end{aligned}$$

A comparison between $\hat{\mathbf{W}}_1$ and $\mathbf{W}_{AS}$ shows that they coincide for $\hat{\mathbf{Q}}$ given by (10).

[Received March 1995. Revised June 1996.]

REFERENCES

Alexander, C. H. (1987), "A Class of Methods for Using Person Controls in Household Weighting," *Survey Methodology*, 13, 183–198.

Bethlehem, J. G., and Keller, W. J. (1987), "Linear Weighting of Sample Survey Data," *Journal of Official Statistics*, 3, 141–153.

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

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Heerschop, M. J., and Lieftinck-Koeijers, C. A. J. (1991), "Registered Unemployment in The Netherlands: Estimation of a Dynamic Population Using Retrospective Information," *The Statistician*, 40, 301–314.

Huang, F., and Fuller, W. (1978), "Nonnegative Regression Estimation for Sample Survey Data," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 300–305.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Lemaître, G., and Dufour, J. (1987), "An Integrated Method for Weighting Persons and Families," *Survey Methodology*, 13, 199–207.

Luery, D. (1986), "Weighting Survey Data Under Linear Constraints on the Weights," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 325–330.

Raghunathan, T. E., and Grizzle, J. E. (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90, 54–63.

Rao, C. R. (1973), *Linear Statistical Inferences and Its Applications* (2nd ed.), New York: Wiley.

Robinson, P. M., and Särndal, C. E. (1983), "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling," *Sankhya*, Ser. B, 45, 240–248.

Särndal, C. E., Swensson, B., and Wretman, J. H. (1992), *Model-Assisted Survey Sampling*, New York: Springer-Verlag.

Seber, G. A. F. (1977), *Linear Regression Analysis*. New York: Wiley.

Wright, R. L. (1983), "Finite Population Sampling With Multivariate Auxiliary Information," *Journal of the American Statistical Association*, 78, 879–884.

Zieschang, K. D. (1986), "A Generalized Least Squares Weighting System for the Consumer Expenditure Survey," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 64–71.

——— (1990), "Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey," *Journal of the American Statistical Association*, 85, 986–1001.