Survey Methodology, June 1997
Vol. 23, No. 1, pp. 23-32
Statistics Canada

Variable Selection for Regression Estimation in Finite Populations

PEDRO L.D. NASCIMENTO SILVA and CHRIS J. SKINNER[1]

ABSTRACT

The selection of auxiliary variables is considered for regression estimation in finite populations under a simple random sampling design. This problem is a basic one for model-based and model-assisted survey sampling approaches and is of practical importance when the number of variables available is large. An approach is developed in which a mean squared error estimator is minimised. This approach is compared to alternative approaches using a fixed set of auxiliary variables, a conventional significance test criterion, a condition number reduction approach and a ridge regression approach. The proposed approach is found to perform well in terms of efficiency. It is noted that the variable selection approach affects the properties of standard variance estimators and thus leads to a problem of variance estimation.

KEY WORDS: Auxiliary information; Calibration; Sample surveys; Subset selection; Ridge regression.

1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977, chap. 7). For the basic case when the population mean $\bar{X}$ of a vector of variables x_i is known and simple random sampling is used, the regression estimator of the population mean $\bar{Y}$ of a survey variable y_i takes the form

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})'b \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i respectively, and b is the sample vector of linear regression coefficients of y_i on x_i.

Regression estimation is useful for at least three reasons. First, it is flexible. Any number of population means of continuous or binary variables can, in principle, be incorporated into $\bar{X}$. In particular, poststratification arises as a special case (Särndal, Swensson and Wretman 1992, sec. 7.6). The procedure also extends to handle complex sampling designs. Second, regression estimation has certain optimal efficiency properties. See, for example, Isaki and Fuller (1982, Theorem 3). Third, $\bar{y}_r$ has the "calibration" property that if y_i is one of the variables of x_i so that $\bar{Y}$ is known then $\bar{y}_r = \bar{Y}$ (Deville and Särndal 1992).

In this paper we consider the question of how to select the x variables for use in the regression estimator. This question is of interest for at least two reasons. First, there is simply the practical reason that in some circumstances the number of potential variables in x_i may be very large. For example, in population censuses in a number of countries values of some variables are recorded on a "short form" for all individuals and values of other variables are collected on a "long form" for a sample. The population means of the short form variables together with their squares, cubes, products and so forth will thus be known. Small area identification will also typically be available. Thus the dimension of x_i as a vector containing functions of the short form variables together with dummy variables representing each small area could easily run into the thousands. In such cases, the selection of x variables becomes a practical necessity.

A second reason is more fundamental for a model-assisted or model-based approach to survey sampling. These approaches may be characterised as follows in the context of regression estimation. First a regression model is selected which has "good predictive power", so that the regression estimator will have "good efficiency". Then, either a design-based approach to inference is adopted in the model-assisted approach (Särndal *et al.* 1992) or model-based prediction is employed in the model-based approach. Although the literature on the latter problem of inference is vast, there seems remarkably little formal attention devoted to the former model selection problem. In practice, the most that seems to happen is that the "main" x variables which account for "most of" the sample R^2 are chosen (*cf.* Särndal *et al.* 1992, sec. 7.9.1). However, more theoretical guidance seems needed, especially when a large number of x variables is available.

A further reason for considering the variable selection problem more formally is that it may help clarify the issue of the impact of variable selection on inference. The problem that sample-based selection of estimators may affect the properties of the selected estimator has long been recognized (Hansen and Tepping 1969, App.) but little study seems to have been made of what the effects may be.

In this paper we consider a variable selection approach aimed at minimising the mean squared error of $\bar{y}_r$. First, however, we study the dependence of the mean squared error of $\bar{y}_r$ on the number of x variables in section 2 and then consider alternative estimators of the mean squared error of $\bar{y}_r$

[1] Pedro L.D. Nascimento Silva, IBGE-Departamento de Metodologia, Avenida Chile 500, Rio de Janeiro-RJ, Brasil; and Professor Chris J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

in section 3. Variable selection procedures based on these estimators are then proposed in section 4.

We contrast our variable selection approach with four existing approaches. First, we consider the traditional approach of using a fixed subset of auxiliary variables regardless of the observed sample. Next, we consider a "condition number reduction procedure" inspired by work of Bankier (1990), in which auxiliary variables are discarded in order to reduce the condition number of a certain cross-products matrix of the x variables.

Third, we follow Bardsley and Chambers (1984) and consider a ridge regression approach. This does not involve variable selection but instead addresses the possible problem of multicollinearity in the regression estimator by modifying the estimator, allowing for some calibration error. Both the ridge regression and condition number reduction procedures have the advantage that they do not require specification of a response variable y, because they aim to provide a single set of "calibration" weights to be used for all survey variables. However, they do not guarantee gains in efficiency. Their results are separated by a line from the results for the other procedures in the tables presented in section 6 to indicate that they differ.

Fourth, we consider variable selection following conventional significance test criteria. Our general view is that the objective of variable selection in regression estimation for finite populations is quite different from the objective of parameter estimation or prediction of y values for single observations in classical regression (Miller 1990). However, it seems desirable to treat such an approach as one benchmark for comparison.

In section 5 we consider properties of the regression estimator following variable selection on the basis of estimated variances. Section 6 describes an empirical study carried out to compare our proposed variable selection procedures with the competing procedures described above. This study used data from a test census carried out in the municipality of Limeira, Brasil, as part of the preparation for the 1991 Brazilian Population Census. Section 7 presents our conclusions and some directions for further research.

2. THE DEPENDENCE OF THE VARIANCE OF THE REGRESSION ESTIMATOR ON THE NUMBER OF x VARIABLES

We begin by defining some notation. Let $U = \{1,...,N\}$ denote a finite population of N distinguishable elements and let $s \subset U$ denote a sample of n distinct elements drawn from U according to a simple random sampling without replacement design. Let $x_i = (x_{i1},...,x_{iq})'$ be the $q \times 1$ vector of auxiliary variables associated with the i-th population element. It is assumed that the sample values of $x_i (i \in s)$, together with the population mean vector $\bar{X} = N^{-1}\sum_{i \in U} x_i$ are known. The vector of sample means is denoted $\bar{x} = n^{-1}\sum_{i \in s} x_i$.

Let y_i denote the value of a survey variable y for the i-th population element and suppose the values of y_i are only observed for $i \in s$. The aim is to estimate the population mean $\bar{Y} = N^{-1}\sum_{i \in U} y_i$.

The regression estimator of $\bar{Y}$ is given by $\bar{y}_r$ in equation (1), where $\bar{y} = n^{-1}\sum_{i \in s} y_i$, $b = \hat{S}_x^{-1}\hat{S}_{xy}$, $\hat{S}_x = n^{-1}\sum_{i \in s}(x_i - \bar{x})(x_i - \bar{x})'$, and $\hat{S}_{xy} = n^{-1}\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})$.

This estimator may be motivated by the underlying linear model

$$y_i = \beta_0 + x_i' \beta + \epsilon_i \tag{2}$$

where the ϵ_i are independent disturbances with zero means and common variance σ^2, since we may write $\bar{y}_r = \hat{\beta}_0 + \bar{X}'\hat{\beta}$, where $\hat{\beta}_0 = \bar{y} - \bar{x}'b$ and $\hat{\beta} = b$ are the least squares estimators of β_0 and β, respectively. Under this model the variance of $\bar{y}_r - \bar{Y}$ conditional on the x_i may be written

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}[1 - n/N + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})]. \tag{3}$$

The final term may be interpreted as the effect of estimating β by b. As the number q of x variables increases the residual variance σ^2 may be expected to decrease, but the term $(\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})$ may increase as $\hat{S}_x^{-1}$ becomes more unstable. An alternative way to interpret this term is to write $\bar{y}_r$ as a weighted estimator $\bar{y}_r = n^{-1}\sum_{i \in s} g_i y_i$, where $g_i = 1 + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})$. Then we may write (3) alternatively as

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}(1 - n/N + c_g^2) \tag{4}$$

where c_g is the sample coefficient of variation of the g_i.

To study the expected dependence of c_g^2 on q we now extend the model by supposing that the x_i are independently and identically normally distributed. Noting the independence of $(\bar{x} - \bar{X})$ and $\hat{S}_x$ and also that $E_M(\bar{y}_r - \bar{Y} \mid x_i) = 0$, we obtain the unconditional variance

$$\begin{aligned} &\mathrm{Var}_M(\bar{y}_r - \bar{Y}) \\ &\quad = \sigma^2 n^{-1}\{1 - n/N + \mathrm{tr}[E_M[(\bar{X} - \bar{x})(\bar{X} - \bar{x})']E_M(\hat{S}_x^{-1})]\} \qquad (5) \\ &\quad = \sigma^2 n^{-1}(1 - n/N)[1 + q/(n - q - 2)] \end{aligned}$$

using the fact that $n^{-1}\hat{S}_x^{-1}$ has an inverse Wishart distribution (Mardia, Kent and Bibby 1979, p. 69 and 85). This result holds for large n even without normality, in the sense that $[1 - n/N + c_g^2]/(1 - n/N)[1 + q/(n - q - 2)]$ still converges to 1 as n increases for fixed q (under weak conditions).

Expression (5) makes the dependence on q explicit. As q increases we may expect σ^2 to decrease but $E_M(c_g^2)$ to increase. The reduction of σ^2 may be expected to be small after a few important x variables are included and thus the variance may be expected to start increasing at some point where the number of x variables is a nonnegligible fraction of the sample size.

Results (4) and (5) are based on strong modelling assumptions and hence provided us only with motivation. In the general case $\bar{x} - \bar{X} = O_p(n^{-1/2})$ (under the randomization distribution with standard regularity conditions) so that the last term of (3) is of $O_p(n^{-2})$. A more general second order asymptotic approximation for the design mean squared error of $\bar{y}_r$ when model (2) need not hold may be obtained by generalising Theorem 4.1 of Deng and Wu (1987). Details are given in Silva (1996).

Our aim is to develop a variable selection procedure that minimizes the estimated mean squared error of $\bar{y}_r$, and estimators of this mean squared error are considered next.

3. ESTIMATION OF THE MEAN SQUARED ERROR OF THE MULTIPLE REGRESSION ESTIMATOR

A simple estimator of the mean squared error of $\bar{y}_r$ is obtained by generalizing expression (7.29) of Cochran (1977, p. 195) to the case of several auxiliary variables:

$$v_s = \frac{1-f}{n}\hat{S}_e \tag{6}$$

where $\hat{S}_e = (n-q-1)^{-1}\sum_{i\in s}\hat{e}_i^2$ and $\hat{e}_i = (y_i - \bar{y}) - (x_i - \bar{x})'b$.

This estimator makes no allowance for the $O(n^{-2})$ component of the mean squared error, however. Thus, as a second mean squared error estimator, we generalize the estimator v_d studied in Deng and Wu (1987) to the case of general q. This is a special case of the model-based, bias-robust variance estimator G_2 originally proposed by Royall and Cumberland (1978), for the case where the residual variances in the model (2) are constant. This estimator is given by

$$v_d = \frac{1-f}{n(n-1)}\sum_{i\in s}\alpha_i\hat{e}_i^2 \tag{7}$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f)/\{(1-f)[1-(x_i-\bar{x})'\hat{S}_x^{-1}(x_i-\bar{x})/(n-1)]\}.$$

We originally conjectured that v_d would be second order unbiased, as Deng and Wu (1987, eq. 4.4) show that it is for the case of $q = 1$. However this turns out not to be the case for general $q > 1$, although it may be expected that the bias of v_d is smaller than that of v_s, as indicated by the second order bias expressions for v_s and v_d obtained by Silva (1996).

A difficulty with v_d as a variance estimator is that it does not generalize easily to complex survey designs. Thus we consider as a third variance estimator a modified version of an estimator proposed by Särndal, Swensson and Wretman (1989), defined as:

$$v_g = \frac{1-f}{n(n-q-1)}\sum_{i\in s}g_i^2\hat{e}_i^2. \tag{8}$$

This estimator may be expected to behave similarly to v_d since $\alpha_i = g_i^2 + O_p(n^{-1/2})$. In the terminology of Särndal *et al.* (1992, p. 232), the g_i are the appropriate *g-weights* under simple random sampling if (2) is adopted as the underlying model. Expression (8) differs from the corresponding estimator proposed by Särndal *et al.* (1989, example 4.4) in that we use the denominator $(n-q-1)$ instead of the original $(n-1)$.

4. VARIABLE SELECTION PROCEDURES

We consider two basic variable selection procedures. First, an *all subsets* approach that involves computing one of the mean squared error estimators v_s, v_d or v_g of section 3 for all 2^q possible subsets of the q auxiliary variables (always including the intercept) and choosing that subset corresponding to the smallest mean squared error estimate. This procedure can clearly involve considerable computation if q is large. Thus as a second procedure, we consider a *forward selection* approach which starts with the sample mean as an estimator, then adds that variable which minimizes the mean squared error estimate. The procedure is repeated until the mean squared error estimate starts to increase, at which point the subset of variables which gave the minimum mean squared error estimate is selected.

These procedures may be contrasted with an approach inspired by the work of Bankier and his associates – see Bankier (1990) and Bankier, Rathwell and Majkowski (1992). We call this a *condition number reduction approach.* To describe the approach, first note that the regression estimator in (1) can alternatively be expressed as

$$\bar{y}_r = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \tag{9}$$

where X_s^* is the $n \times (q+1)$ matrix with $x_i^{*\prime} = (1, x_{i1}, \ldots, x_{iq})' = (1 \vdots x_i')$ as its i-th row, $\bar{x}^* = (1 \vdots \bar{x}')'$ and $\bar{X}^* = (1 \vdots \bar{X}')'$ are the sample and population mean vectors of x_i^* respectively, and y_s is the $n \times 1$ vector with the sample observations of the response.

The regression estimator thus depends on the inversion of the cross-products matrix $X_s^{*\prime}X_s^*$, a matrix which can sometimes become ill-conditioned and thereby inflate the variance of the regression estimator.

Bankier (1990) proposed a two-step procedure for computing regression estimators of means (or totals) in which columns of the auxiliary data matrix X_s^* were eliminated in order to reduce the condition number of the cross-products matrix $X_s^{*\prime}X_s^*$, as well as to avoid undesirable situations (negative or outlying weights, rare characteristics, or exact linear dependence between columns). Bankier *et al.* (1992) describe in detail the procedure as applied to the 1991 Canadian Population Census. It is worth noting that the approach developed by Bankier and associates, although incorporating variable selection, is not targeted at achieving efficiency for a particular survey variable. Its main focus is on calibration, while at the same time providing a single set of weights that are used for all survey variables.

The condition number reduction approach that we consider can be described by the algorithm below, which adopts a backward elimination procedure to discard auxiliary variables generating large condition numbers for the cross-products matrix $CP = X_s^{*\prime} X_s^{*}$, instead of the forward inclusion of variables described by Bankier *et al.* (1992).

1) Compute the cross-products matrix $CP = X_s^{*\prime} X_s^{*}$ considering all the columns initially available (saturated subset).

2) Compute the Hermite canonical form of CP, say H (see Rao 1973, p.18), and check for singularity by looking at the diagonal elements of H. Any zero diagonal elements in H indicate that the corresponding columns of $X_s^{*\prime} X_s^{*}$ (and X_s^{*}) are linearly dependent on other columns (see Rao 1973, p. 27). Each of these columns is eliminated by deleting the corresponding rows and columns from $X_s^{*\prime} X_s^{*}$.

3) After removing any linearly dependent columns, the condition number $c = \lambda_{\max}/\lambda_{\min}$ of the reduced CP matrix is computed, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest of the eigenvalues of CP, respectively. If $c < L$, a specified value, stop and use all the auxiliary variables remaining.

4) Otherwise perform backward elimination as follows. For every k, drop the k-th row and column from CP, and recompute the eigenvalues and the condition number of the reduced matrix. Compute the condition number reductions $r_k = c - c_k$, where c_k is the condition number after dropping the k-th row and column from CP. Determine $r_{\max} = \max_k (r_k)$and $k_{\max} = \{k : r_{\max} = r_k\}$ and eliminate the column $k_{\max}$ by deleting the $k_{\max}$ row and column from CP. Make $c = c_{k_{\max}}$ and iterate while $c \geq L$ and $q \geq 2$, starting each new iteration with the reduced CP matrix resulting from the previous one.

One further approach that we consider is the ridge regression estimator of Bardsley and Chambers (1984). It does not rely on selecting subsets from the auxiliary variables available, but rather on relaxing the calibration properties of the regression estimator in favour of more stable estimates. The ridge regression estimator is given by

$$\bar{y}_{BC} = [n\bar{y} + (N\bar{X}^{*} - n\bar{x}^{*})'(\lambda C^{-1} + X_s^{*\prime} X_s^{*})^{-1} X_s^{*\prime} y_s]/N \qquad (10)$$

where λ is a scalar ridging parameter and C is a diagonal matrix of "cost" coefficients associated with the calibration errors tolerated when estimating totals of the auxiliary variables using $\bar{y}_{BC}$.

Bardsley and Chambers (1984) suggested that the specification of the matrix C could be used to control the influence of each auxiliary variable on the resulting estimator of the response mean, thus imitating the subset selection process. As for the ridging parameter λ, they suggested taking the smallest value such that all the implicit case weights are not smaller than $1/N$ (or 1 for estimating totals).

5. PROPERTIES OF REGRESSION ESTIMATORS AFTER VARIABLE SELECTION

For our basic variable selection procedures, a set of estimation strategies $\boldsymbol{S} = \{(\bar{y}_r^{\gamma}, v^{\gamma}); \gamma \in \Gamma\}$ is considered, where $\bar{y}_r^{\gamma}$ and v^{γ} are the regression estimator and an estimator of its variance respectively for a subset γ of the q auxiliary variables available, and Γ is the set of all subsets. The variable selection procedure selects a subset γ^* from Γ according to a rule which is determined by the data and by $\boldsymbol{S}$, and the resulting point estimator is $\bar{y}_r^{\gamma^*}$.

For each fixed subset γ, it follows under standard regularity conditions (Isaki and Fuller 1982) that $\bar{y}_r^{\gamma}$ is consistent for the population mean $\bar{Y}$, that is $\bar{y}_r^{\gamma} - \bar{Y} = o_p(1)$. Now, for given $\delta > 0$, $|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta$ implies $|\bar{y}_r^{\gamma} - \bar{Y}| > \delta$ for some γ, and so we have

$$\Pr(|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta) \leq \sum_{\gamma \in \Gamma} \Pr(|\bar{y}_r^{\gamma} - \bar{Y}| > \delta) \qquad (11)$$

and because Γ is finite, the right hand side of (11) converges to zero, and it follows that $\bar{y}_r^{\gamma^*}$ is also consistent.

The distribution of $\bar{y}_r^{\gamma^*}$ will, however, depend on the selection rule in a complex way. See Grimes and Sukhatme (1980) for an investigation of the efficiency of $\bar{y}_r^{\gamma^*}$ in the simplest case when there are just two possible estimators: a regression estimator with one x variable and a difference estimator (a special case of which is the mean) and the variables are jointly normally distributed.

In contrast to the consistency of $\bar{y}_r^{\gamma^*}$, there is no reason why v^{γ^*} should be consistent for $\mathrm{Var}(\bar{y}_r^{\gamma^*})$, even if v^{γ} is consistent for $\mathrm{Var}(\bar{y}_r^{\gamma})$ for each fixed γ. In particular we may expect v^{γ^*} to underestimate $\mathrm{Var}(\bar{y}_r^{\gamma^*})$ if the selection rule is such that v^{γ^*} is the minimum of the v^{γ}. This effect is similar to the well known overestimation of R^2 after subset selection in standard multiple linear regression (Miller 1990, p. 7-10).

6. A SIMULATION STUDY

In this section we present a small simulation study carried out to evaluate the performance of the alternative variable selection procedures considered. We took as our simulation population a data set comprising 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil.

This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. The test consisted of two rounds of data collection. In the first round, each enumerator would visit all the occupied households in a given enumeration area (an area with between 200 and 300 households on average) and would fill in a short questionnaire. This form contained a few questions about characteristics of the household and about each member of the household (sex, age, relationship to head of household and literacy). For heads of household only, a question on education and another about monthly total income were also included. The reported monthly total income for heads of household provides only a proxy to the actual income, due to the limitations of the interviewing process in this first round of data collection.

Then a second round of data collection was undertaken in each enumeration area. The same enumerators would visit a sample of 1 in 10 of the households (selected systematically from the list of occupied households compiled in the first round of data collection) to obtain information using a long (more detailed) questionnaire, which contained all the questions asked in the short form plus many other questions.

The size of the surveyed population was approximately 44,000 households with 188,000 individuals. The sample size was roughly 10% of the population size. For reasons of computational cost, we used in our simulation study a subpopulation comprising all the sample records for 426 heads of household living in 20 of the 170 enumeration areas. We chose these records as our simulation population because they contain all the detailed information provided in the sample questionnaire, as well as the proxy information available from the first round interviews using the short form.

We considered total monthly income, as obtained from the long form, as the main response variable (y) together with 11 potential auxiliary variables, namely:

x_1 = indicator of sex of head of household equal male;
x_2 = indicator of age of head of household less than or equal to 35;
x_3 = indicator of age of head of household greater than 35 and less than or equal to 55;
x_4 = total number of rooms in household;
x_5 = total number of bathrooms in household;
x_6 = indicator of ownership of household;
x_7 = indicator that household type is house;
x_8 = indicator of ownership of at least one car in household;
x_9 = indicator of ownership of colour TV in household;
x_{10} = years of study of head of household;
x_{11} = proxy of total monthly income of head of household.

From these 11 variables, we constructed two alternative sets of auxiliary variables for our simulations. The first set was defined by taking five auxiliary variables, namely $x_1,...,x_4$ and x_{11}, that have reasonable explanatory power in predicting y, especially due to the presence of the proxy income x_{11}. The second set we considered contained ten auxiliary variables, namely $x_1,...,x_{10}$, which due to the exclusion of x_{11}, has smaller predictive power than the previous one. For reference, the population correlation matrix for the survey variable y and the 11 auxiliary variables in the population is given in Table 3.

We then selected 1,000 samples of size 100 from this simulation population by simple random sampling without replacement.

Before proceeding to examine the detailed simulation results, we first consider the potential for gains from variable selection following the motivating model-based discussion of section 2. Recall from equation (4) that under model (2) the conditional variance of $\bar{y}_r$ is inflated by a term c_g^2 because of estimation of β. We evaluated the distribution of c_g^2 over the 1,000 samples for both the cases of five and ten auxiliary variables. For the case of five auxiliary variables, the median value of c_g^2 was 0.036, with upper quartile of 0.056 and maximum 0.255. This accords roughly with equation (5) which implies that under the model the expected value of c_g^2 is $(1 - n/N)q/(n - q - 2) = 0.041$. Note that the wide variation of c_g^2 across samples suggests that it may be sensible to adopt a procedure which selects a different set of variables for each sample. The variation of c_g^2 is even greater for the case of ten auxiliary variables, when the median was 0.078, the upper quartile was 0.107 and the maximum was 0.329, which also accords roughly with the expected value under the model of 0.087, according to equation (5). This interpretation clearly depends on the validity of the model (2), which is doubtful for these data, but it does suggest that there are potential efficiency gains to be made from variable selection.

Another way to assess the potential for efficiency gains from variable selection is to compute approximations to the variance of the regression estimator considering various subsets of the auxiliary variables available, using all the population records. Figure 1 displays a plot of the approximation given by a finite population version of equation (5) computed for increasing subsets of the ten auxiliary variables, where the variable added at each step is the one yielding the biggest decrease in the approximation. The values of the standard first order design-based approximation $(1 - f)S_e/n$ are also plotted for reference, although as has already been noted, this approximation is monotone non-increasing when new auxiliary variables are added. Simulation estimates of the mean squared error for the regression estimator corresponding to each subset are also plotted. The plot shows clearly that if a standard regression estimator with a fixed set of auxiliary variables is to be used, the subset with five predictors would be the best choice when

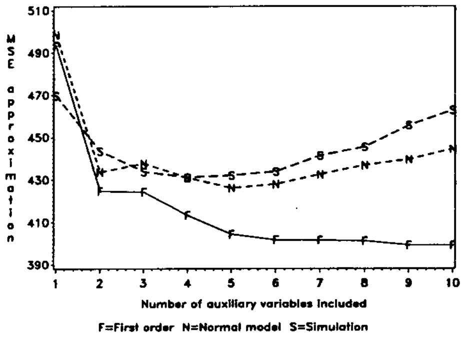

Figure 1. Finite population approximations and simulation estimations for the MSSE of the regression estimator computed for increasing subsets of the ten auxiliary variables.

the normal approximation for the variance based on expression (5) was considered, whereas the saturated subset would be chosen in case the standard design-based approximation for the variance was considered. The plot also reveals that the simulation estimates of the mean squared error agree more closely with the normal model approximation than with the standard first order approximation, especially for larger subsets of auxiliary variables. Similar results are achieved when corresponding variance approximations are computed given the set of five auxiliary variables.

Hence both the simulation distributions of c_g^2 and the finite population approximations to the variance of the regression estimator indicate that there are potential efficiency gains to be made from variable selection for this population. To investigate this for our data we now proceed to describe the details of the simulation study.

For each sample replicate (say s) and for each of the two alternative sets of auxiliary variables considered, estimates of the population mean of total monthly income were computed, as well as corresponding variance estimates, using a number of estimation strategies. Each estimation strategy is defined as a combination of a subset selection procedure, an estimator for the mean and a corresponding variance estimator. The list of all strategies considered follows.

SM) Sample mean estimator, with no auxiliary variables $(\bar{y}, v_s)$. This strategy provides the standard against which all the others will be compared.

Fs) Forward selection of auxiliary variables with $(\bar{y}_r, v_s)$.

Fd) Forward selection of auxiliary variables with $(\bar{y}_r, v_d)$.

Fg) Forward selection of auxiliary variables with $(\bar{y}_r, v_g)$.

Bs) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_s)$.

Bd) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_d)$.

Bg) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_g)$.

FI) Fixed subset of auxiliary variables with $(\bar{y}_r, v_s)$.

SS) Saturated subset of auxiliary variables with $(\bar{y}_r, v_s)$.

FR) Forward subset selection using SAS PROC REG, with $(\bar{y}_r, v_s)$.

CN) Condition number reduction subset selection procedure with $(\bar{y}_r, v_s)$.

RI) Ridge regression estimator with saturated subset of auxiliary variables and a variance estimator that we denote v_{DC}, proposed by Dunstan and Chambers (1986), $(\bar{y}_{BC}, v_{DC})$.

Strategies Fs to Bg are variations of the two procedures we proposed for subset selection arising from the use of the three mean squared error estimators considered in section 3. Strategies FI and SS use the same set of auxiliary variables irrespective of the sample selected. In SS the saturated subset including all auxiliary variables available is always used. In FI a subset was chosen from each of the two sets with five (x_1, x_4, x_{11} chosen) or ten ($x_1, x_2, x_5, x_8, x_{10}$ chosen) auxiliary variables considered, by applying a standard forward subset selection regression procedure to the population dataset. The selected subsets were then used for every sample, thus the name "fixed subset" strategy for FI. This strategy would not be feasible in practice because the population information would not be available for the response, but it was considered as a theoretical "best possible scenario" under the traditional approach.

For the strategy FR, SAS PROC REG was used "naively" to perform a standard forward subset selection for each sample. The *p-value* used to decide whether a new variable should be included was the default of the procedure, namely 0.50. For more details, see SAS (1990, p. 1397).

For the condition number reduction subset selection strategy CN, the value used for the parameter L that controls the method was 1,000. For the ridge regression estimator strategy RI, the cost coefficients associated with calibration errors for different variables were all set equal to 1. After having chosen the value of λ that guarantees all the weights are not less than $1/N$, the weights were rescaled such that they sum to exactly 1, in order to ensure exact calibration when estimating the population size.

For any estimation strategy, the estimates of the population mean and its mean squared error for the sample s are denoted by $\bar{y}(s)$ and $v[\bar{y}(s)]$ respectively. The simulation results for each estimation strategy were summarised by computing estimates of the bias, mean squared error (MSE), and average of mean squared error estimates (AVMSE) from the set of 1,000 sample replicates, given respectively by

$$\text{BIAS} = \sum_s [\bar{y}(s) - \bar{Y}]/1{,}000 \qquad (12)$$

$$\text{MSE} = \sum_s [\bar{y}(s) - \bar{Y}]^2/1{,}000 \qquad (13)$$

$$\text{AVMSE} = \sum_s v[\bar{y}(s)]/1{,}000. \qquad (14)$$

A measure of efficiency was also calculated for each strategy by dividing the corresponding simulation mean squared error by the simulation mean squared error for the sample mean (strategy SM) and multiplying the result by 100. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each estimation strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 2.

Table 1 displays the simulation results for estimation of the population mean of the response variable given the set of five auxiliary variables ($x_1 - x_4, x_{11}$) with larger predictive power. In this case, the use of the regression estimator greatly improves precision for every estimation strategy employed, except for subset selection using condition number reduction (CN). The bias was negligible (less than 1% in terms of the absolute relative bias) for all estimation strategies (the population mean of y is 194.34) except perhaps RI, which displayed a slight bias.

Table 1
Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Five Auxiliary Variables (x_1 - x_4, x_{11}) Available

Estimation strategy	BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM) Sample mean ($\bar{y}, v_s$)	0.25	620.09	619.05	100.00	91.8
Fs) Forward ($\bar{y}_r, v_s$)	0.40	233.78	239.62	37.70	82.7
Fd) Forward ($\bar{y}_r, v_d$)	-1.25	188.08	196.88	30.33	82.0
Fg) Forward ($\bar{y}_r, v_g$)	-1.28	188.38	192.73	30.38	81.1
Bs) Best ($\bar{y}_r, v_s$)	0.44	236.90	239.49	38.20	82.7
Bd) Best ($\bar{y}_r, v_d$)	-1.22	190.52	196.84	30.72	82.0
Bg) Best ($\bar{y}_r, v_g$)	-1.24	190.83	192.71	30.77	81.1
FI) Fixed ($\bar{y}_r, v_s$)	0.29	227.90	241.24	36.75	83.3
SS) Saturated ($\bar{y}_r, v_s$)	0.30	233.58	242.32	37.67	82.5
FR) PROC REG ($\bar{y}_r, v_s$)	0.38	235.86	240.26	38.04	82.5
CN) Cond. num. red. ($\bar{y}_r, v_s$)	0.34	507.33	483.63	81.82	89.8
RI) Ridge ($\bar{y}_{BC}, v_{DC}$)	2.12	304.95	257.07	49.18	82.5

[1] Nominal 95% coverage.

There was no difference between the results for strategies based on forward selection (Fs-Fg) and corresponding strategies based on selection from all possible subsets (Bs-Bg). Hence the faster and cheaper forward selection procedures are preferable.

Amongst the strategies using forward subset selection, Fd and Fg (with v_d and v_g as the mean squared error estimators respectively) yielded greater efficiency, and performed very similarly. Note also that Fd and Fg performed better than FI and SS, the strategies that adopted the regression estimator with a fixed subset of the five auxiliary variables for every sample. This is true both for the saturated subset (SS) and when the fixed subset was chosen using information from the whole population (FI). This shows that one can do better than the traditional approach of using the regression estimator with a fixed set of auxiliary variables, by using an adaptive procedure that chooses the "best" regression estimator (subset) for each given sample, at least when the target response variable is the one considered for subset selection. This property was suggested by the wide variation in the values of c_g^2 between samples, where we may expect to benefit from a strategy which selects fewer x variables for samples with the largest values of c_g^2.

Comparison with the adaptive strategy FR, which used the standard subset selection available in PROC REG of SAS, shows that a criterion using an appropriate estimator of the mean squared error of the regression estimator makes some difference. FR yielded similar efficiency to that of traditional fixed subset strategies (FI-SS).

A more striking result is the low efficiency achieved by the subset selection procedure based on condition number reduction (CN) compared to all the other strategies based on the regression estimator. This was not unexpected, because that procedure did not take the response variable into account. This favours the argument that when the mean of some specified response variable is the main target for inference, this should be taken into account when selecting the auxiliary variables to use in connection with the regression estimator.

When the set of five auxiliary variables was considered, we also observed that, for every sample, the first variable eliminated to reduce the condition number was proxy income (x_{11}). This happened because eigenvalues (and hence condition numbers) of the CP matrix are dependent on the units of measurement of the auxiliary variables. Because all other auxiliary variables are counts of some kind, proxy income is the variable with the largest variance by far. Its exclusion for every sample provides some explanation for the poor performance of this approach, because it is the best single predictor for the response.

This difficulty was not apparent in Bankier's work, because in the target application of his procedure, the sample data from the 1991 Canadian Population Census, all the auxiliary variables considered were counts of persons, families or households, thus measured in similar units.

Unlike the eigenvalues of the CP matrix, the regression estimator is invariant to location and scale transformation of the auxiliary variables. To remove the arbitrary dependence of the condition number approach on the units of the auxiliary variables, it is therefore natural to standardise these variables first and to compute the condition number of the sample correlation matrix $\hat{R}_x$ rather than $X_s^{*\prime}X_s^*$. However this was tried and even modest values of L (100) failed to cause elimination of any auxiliary variables, which resulted in the saturated set being used every time, so that CN reduced to SS.

The strategy based on the ridge regression estimator (RI) performed worse than the saturated subset strategy (SS) in terms of efficiency. It also displayed some bias for estimating the mean squared error. This loss of efficiency is due to the requirement that all the weights should be greater than or equal to $1/N$, which was imposed only under this strategy. On the other hand, it performed much better than the condition number reduction strategy CN in terms of efficiency.

In terms of the empirical coverage rates, only the condition number reduction strategy CN performed close to SM (sample mean), both leading to modest undercoverage. All the other strategies based on regression estimation yielded similar coverage rates, well below the target of 95%.

Results for the simulation carried out with the set of ten auxiliary variables (x_1 - x_{10}) are displayed in Table 2 below. As expected, these results show that the strategies that use the regression estimator still provide some gain in efficiency over the sample mean. However these gains are not as large as those reported in Table 1, when there are five auxiliary variables with higher explanatory power. As before, adaptive strategies based on forward subset selection performed similarly to their counterparts based on best subset selection from all possible subsets. Adaptive strategies using v_d or v_g as the estimator of the mean squared error were again slightly more efficient than the corresponding strategies based on v_s, although in this case at the expense of larger undercoverage of the corresponding nominal 95% confidence intervals.

The more efficient adaptive estimation strategies (Fd, Fg, Bd and Bg) display nonnegligible bias for both the population mean and for the mean squared error. In contrast, strategies FI and SS present no significant bias for the mean, although there is some bias in the mean squared error estimation under strategy SS. Note particularly the large negative bias of the estimators of the mean squared error, as indicated by the differences between the columns labelled MSE and AVMSE in Table 2. This appears to be worse for strategies Fd, Fg, Bd and Bg, followed by Fs and Bs, and not so bad for SS, FR and CN.

Comparing Fd and Fg with CN, there is a moderate gain in efficiency over the condition number reduction procedure, at the expense of some increased bias in both the mean and mean squared error estimators. Thus, even when the predictive power of the available auxiliary variables is not large, it is still possible to gain efficiency over strategy CN.

A bad choice of fixed subset (as for example, the saturated subset used in strategy SS) could yield poor results in terms of efficiency and also some bias in the mean squared error estimation. However, if for example v_d was used as the estimator for the mean squared error under strategy SS instead of v_s, there would be no apparent bias (the AVMSE observed in that case was 459.67, hence much closer to the estimated simulation mean squared error of 462.71).

The ridge regression estimator was again slightly inferior to the saturated subset strategy (SS), but now without any apparent bias in estimating the mean or the mean squared error. It outperformed the condition number reduction strategy CN once again in terms of efficiency, albeit by a smaller margin. It also performed well in terms of empirical coverage.

Strategy FR performed similarly to the fixed subset strategies FI and SS again, and so was outperformed by strategies using a specialized criterion based on an estimator of the mean squared error of the regression estimator such as v_d or v_g.

These results suggest that, when estimating the population mean of a single response, the proposed adaptive procedures combining the regression estimator with some form of subset selection based on an appropriate mean squared error estimaator can offer some useful improvements in efficiency against its competitors. However such strategies may introduce some bias when the predictive power of the auxiliary variables available is not large, and the corresponding MSE estimators may be substantially biased, leading to poor coverage.

Table 2

Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Ten Auxiliary Variables (x_1-x_{10}) Available

Estimation strategy		BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM)	Sample mean ($\bar{y}, v_s$)	0.25	620.09	619.05	100.00	91.8
Fs)	Forward ($\bar{y}_r, v_s$)	0.06	468.46	397.99	75.55	86.7
Fd)	Forward ($\bar{y}_r, v_d$)	-8.12	434.27	338.90	70.03	81.7
Fg)	Forward ($\bar{y}_r, v_g$)	-7.90	433.71	328.46	69.94	81.6
Bs)	Best ($\bar{y}_r, v_s$)	-0.00	466.16	397.59	75.18	86.6
Bd)	Best ($\bar{y}_r, v_d$)	-7.90	434.54	336.88	70.08	81.5
Bg)	Best ($\bar{y}_r, v_g$)	-7.60	433.26	326.05	69.87	81.6
FI)	Fixed ($\bar{y}_r, v_s$)	0.45	490.49	461.86	79.10	89.0
SS)	Saturated ($\bar{y}_r, v_s$)	-0.20	462.71	413.17	74.62	86.9
FR)	PROC REG ($\bar{y}_r, v_s$)	-0.07	466.13	399.34	75.17	86.4
CN)	Cond. num. red. ($\bar{y}_r, v_s$)	3.49	562.91	450.36	90.78	87.3
RI)	Ridge ($\bar{y}_{BC}, v_{DC}$)	1.05	480.18	472.82	77.44	89.4

[1] Nominal 95% coverage.

Table 3
Correlation Matrix for Variables Used in the Simulation Study with the 1988 Census Population

Variable	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	0.23										
x_2	-0.04	0.20									
x_3	0.17	0.07	-0.40								
x_4	0.47	0.13	-0.15	0.12							
x_5	0.48	0.09	-0.11	0.15	0.83						
x_6	0.05	-0.09	-0.32	-0.03	0.22	0.20					
x_7	-0.17	0.01	-0.12	-0.01	-0.17	-0.31	0.16				
x_8	0.38	0.29	0.07	0.17	0.44	0.41	0.13	-0.20			
x_9	0.20	0.08	-0.06	0.04	0.30	0.25	0.16	-0.13	0.37		
x_{10}	0.43	0.23	0.33	0.17	0.39	0.39	-0.10	-0.30	0.49	0.26	
x_{11}	0.78	0.23	-0.00	0.22	0.54	0.54	0.01	-0.19	0.41	0.21	0.49

7. CONCLUSIONS AND FUTURE DIRECTIONS

Our results suggest that, when using regression estimation, there is potential for some gain in efficiency by adopting a variable selection procedure based on one of the mean squared error estimators v_d or v_g. Under SRS, and considering the limited simulation evidence, there seems little to choose between these two mean squared error estimators.

Forward subset selection procedures were as effective as those based on searches carried out considering all possible subsets, which involve much more computation. Our results also indicate that it is possible to improve over subset selection procedures based on condition number reduction whenever a specific response variable is of interest.

One problem with a variable selection approach is that the associated variance estimation is likely to become biased for the estimation of the overall mean squared error of the regression estimator following variable selection, thus leading to poor coverage of standard confidence interval procedures. Further research is necessary to investigate possible alternative variance estimation procedures.

This paper has focused on the use of regression estimation to reduce sampling variance in the classical sampling context. In practice, regression estimation is widely used to correct for biases arising from non-sampling errors. In such applications the question of how many auxiliary variables to use is also an important one. Some variables might be included for reasons unrelated to sampling error, for example because they are known to be important determinants of nonresponse. Nevertheless, as the number of auxiliary variables increases the sampling variance may also eventually increase and we suggest that a decision rule to limit the number of auxiliary variables employed might still usefully be based on sampling variance considerations. In the presence of nonsampling bias, the difference between $\bar{x}$ and $\bar{X}$ will generally be of $O_p(1)$ not $O_p(n^{-1/2})$ and so the results of this paper are not directly applicable. Further research is therefore needed to consider the extension of our approach to this case.

Further research is also necessary to extend our approach to complex sampling designs. One possible approach for the general regression estimators, considered *e.g.* by Särndal *et al.* (1992, sec. 6.4), would be to replace the weights g_i by the "generalized" weights, described by Särndal *et al.* (1992, eq. 6.5.9), and to base variable selection on the minimization of the generalized version of v_g given by Särndal *et al.* (1992, eq. 6.6.4).

ACKNOWLEDGEMENTS

Pedro L.D. Nascimento Silva is grateful to CVCP-UK, CNPq-Brasil and IBGE-Brasil for financial support. The authors are grateful to Ray Chambers, Danny Pfeffermann, Jon Rao, Michael Bankier and two anonymous referees for comments. Michael Bankier was also very helpful for providing documentation and software about his GLSEP procedure.

REFERENCES

BANKIER, M.D. (1990). Two Step Generalized Least Squares Estimation. Ottawa: Statistics Canada, Social Survey Methods Division, internal report.

BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Methodology Branch Working Paper, SSMD, 92-007E, Statistics Canada.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

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

DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in mltipurpose surveys. *Applied Statistics*, 35, 276-280.

GRIMES, J.E., and SUKHATME, B.V. (1980). A regression-type estimator based on preliminary test of significance. *Journal of the American Statistical Association*, 75, 957-962.

HANSEN, M.H., and TEPPING, B.J. (1969). Progress and problems in survey methods and theory illustrated by the work of the United States Bureau of the Census. *New Developments in Survey Sampling*, (N.L. Johnson and H. Smith Jr., Eds.). New York: John Wiley & Sons.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MILLER, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). New York: John Wiley & Sons.

ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide* (Version 6, Vol. 2, 4th ed.). Cary, NC: SAS Institute Inc.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SILVA, P.L.D.N. (1996). Some Asymptotic Results on the Mean Squared Error of the Regression Estimator Under Simple Random Sampling Without Replacement. Southampton: University of Southampton, Centre for Survey Data Analysis Technical Report 96-2.