VARIABLE SELECTION FOR REGRESSION ESTIMATION IN THE PRESENCE OF NONRESPONSE

C J Skinner, University of Southampton and P L d Nascimento Silva, Instituto Brasileiro de Geografia e Estatistica
C J Skinner, Department of Social Statistics, University of Southampton, Southampton, S017 1BJ UK.

Key words: Auxiliary Information, Finite Population, Survey Sampling.

1. Introduction

Nascimento Silva and Skinner (1997) (hereafter NS) consider the selection of auxiliary variables in the regression estimation of finite population means under simple random sampling. They consider the classical objective of regression estimation, which is to improve precision compared to the sample mean, and note that the variance of the regression estimator is not necessarily minimised by including all possible auxiliary variables because of the effect of estimating regression coefficients. They consider alternative approaches to selecting auxiliary variables based on the minimisation of various estimators of the variance of the regression estimator.

In this paper we extend the approach of NS to handle nonresponse. The primary new issue is that of bias. Under simple random sampling, the regression estimator is approximately unbiased (under a design-based approach) for any choice of auxiliary variables. Hence NS base their choice of auxiliary variables only on variance considerations. Under nonresponse, however, the estimator may be biased.

We shall assume in this paper that there is no bias if the maximal choice of auxiliary variables is used, but that bias may arise with subsets of auxiliary variables. We propose to select auxiliary variables according to the estimated mean squared error (MSE) of the regression estimator, thus trading off nonresponse bias against variance.

The use of regression estimation to compensate for nonresponse has been considered widely in the literature. See, for example, Cassel, Särndal and Wretman (1983), Särndal and Swensson (1987), Bethlehem (1988) and Fuller, Loughlin and Baker (1994). Our paper extends an approach in the latter paper, which compares the estimated MSE of the regression estimator to that of the estimator which makes no use of the auxiliary information. Our approach is also related to ideas for weighting class methods which suggest choosing weighting classes according to MSE considerations (Tremblay, 1986; Kalton and Maligalig, 1991).

2. Basic Theory

Let r denote the set of respondents, which is a subset of a finite population U. Let y be the survey variable of interest, with values y_i observed for $i\epsilon r$ and suppose the aim is to estimate the population mean $\bar{Y} = N^{-1}\sum_U y_i$, where N is the size of U. Auxiliary information is assumed available on the population means of the survey variables $x_{(1)},...,x_{(J)}$, that is we suppose values $x_{i1},...,x_{iJ}$ of these variables are observed for $i\epsilon r$ and that $\bar{X}_{(j)} = N^{-1}\sum_{i\epsilon U} x_{ij}$ is known for j = 1,...,J.

Let A denote the subset of the auxiliary variables $\{x_{(1)},..., x_{(J)}\}$ to be used in regression estimation and B the complementary subset of variables not used. Write $x_i = (x_{iA}\ x_{iB})$ as the $1\times J$ vector, containing the values $x_{i1},...,x_{iJ}$ ordered such that x_{iA} is a $1\times J_A$ vector containing values of the selected auxiliary variables and x_{iB} is a $1\times J_B$ vector ($J_A + J_B = J$) containing values of the remaining variables ($i\epsilon U$). Let $\bar{X} = (\bar{X}_A \bar{X}_B)$ be the corresponding vector based on the $\bar{X}_{(j)}$ and let $X = (X_A\ X_B)$ be the corresponding $n_r\times J$ matrix with rows x_i, $i\epsilon r$, where n_r is the size of r. Let Y be the $n_r\times 1$ vector of values y_i ($i\epsilon r$).

We define the regression estimator of $\bar{Y}$ based upon the information on the selected auxiliary variables as

$$\bar{y}_{reg,A} = \bar{X}_A \hat{\beta}_A, \qquad (1)$$

where

$$\beta_A = (\sum_r w_i x_{iA} x_{iA}^{/})^{-1} \sum_r w_i x_{iA} y_i$$

$$= (X_A^{/} W X_A)^{-1} X_A^{/} W Y,$$

w_i is a given weight, $W = \text{diag}(w_i)$ and $X_A^{/} W X_A$ is assumed non-singular. We also assume that x_{iA} always includes unity as its first element. The weights w_i are treated as arbitrary but fixed. They may, for example, be of the form $w_i \propto \pi_i^{-1}$, where π_i is the probability of sampling unit i. For the case where all auxiliary variables are selected and $x_i = x_{iA}$, we write

$$\bar{y}_{reg} = \bar{X}\beta \qquad (2)$$

and refer to this estimator as the saturated regression estimator.

We wish to estimate the mean squared error (MSE) of $\bar{y}_{reg,A}$ for alternative choices of subset A. We shall evaluate the MSE and its components with respect to the joint distribution induced by the sampling scheme, the nonresponse mechanism and the model generating the y_i values. We may write $\bar{y}_{reg,A}$ as a linear estimator in the form

$$\bar{y}_{reg,A} = n_r^{-1} \sum_r g_{Ai} y_i \qquad (3)$$

where $g_{Ai} = \bar{X}_A (X_A^{/} W X_A)^{-1} w_i x_{iA}$. (4)

The estimation of the variance of $\bar{y}_{reg,A}$ has been widely discussed (for example, Särndal et al, 1992, sect 6.6) and will not be pursued here. To take a specific simple choice of variance estimator, we shall use

$$v(\bar{y}_{reg,A}) = \sum_r g_{Ai}^2 \hat{e}_{Ai}^2 / [n_r(n_r - J_A)], \qquad (5)$$

where $\hat{e}_{Ai} = y_i - x_{Ai}\beta_A$, following NS but ignoring the finite population correction. This estimator is natural when the observations may be treated as independent and identically distributed. Fuller et al (1994) consider a similar estimator for the more general case of stratified two-stage sampling.

In order to estimate the bias of $\bar{y}_{reg,A}$, we shall suppose that the saturated regression estimator is unbiased to first order of approximation. Fuller et al (1994) discuss alternative sufficient conditions for this to hold. In particular, $\bar{y}_{reg}$ is unbiased if the following linear model holds

$$y_i = x_i\beta + e_i,\ E(e_i \mid x_i) = 0, \qquad (6)$$

and the combined nonresponse and sampling mechanism is ignorable given the x_i, that is the selection of the respondents r is conditionally independent of the y_i given the x_i for $i \epsilon U$. Under the assumption that $\bar{y}_{reg}$ is unbiased we estimate the bias of $\bar{y}_{reg,A}$ by

$$\hat{B}(\bar{y}_{reg,A}) = \bar{y}_{reg,A} - \bar{y}_{reg} \qquad (7)$$

and estimate the MSE of $\bar{y}_{reg,A}$ by

$$\hat{MSE}(\bar{y}_{reg,A}) = v(\bar{y}_{reg,A}) + \hat{B}(\bar{y}_{reg,A})^2 - v[\hat{B}(\bar{y}_{reg,A})], \qquad (8)$$

where $v[\hat{B}(\bar{y}_{reg,A})]$ is an estimator of the variance of $\hat{B} = \hat{B}(\bar{y}_{reg,A})$ to be derived. A possible modification, following Fuller et al (1994), would be to replace $\hat{B}^2 - v(\hat{B})$ in (8) by zero, if it is negative.

Before proceeding to consider the estimation of the variance of $\hat{B}$, we first present a model-based argument to justify the choice of (7) as the estimator of the bias of $\bar{y}_{reg,A}$. Let

$$X_{B|A} = X_B - X_A (X_A^{/} W X_A)^{-1} X_A^{/} W X_B, \qquad (9)$$

be the residual part of X_B obtained by subtracting its projection, weighted by W, on X_A. Correspondingly, let

$$\bar{X}_{B|A} = \bar{X}_B - \bar{X}_A (X_A^{/} W X_A)^{-1} X_A^{/} W X_B.$$

Since a regression estimator is invariant to a linear transformation of the auxiliary variables and X_A and $X_{B|A}$ are orthogonal ($X_A' W X_{B|A} = 0$), the saturated regression estimator $\bar{y}_{reg}$ may be expressed as

$$\bar{y}_{reg} = \bar{X}_A (X_A' W X_A)^{-1} X_A' W Y + \bar{X}_{B|A} (X_{B|A}' W X_{B|A})^{-1} X_{B|A}' W Y . \quad (10)$$

Thus we may write

$$\bar{y}_{reg} = \bar{y}_{reg,A} + \bar{X}_{B|A} \beta_{B|A} = \bar{y}_{reg,A} + \bar{y}_{reg,B|A}, \quad (11)$$

where

$$\bar{y}_{reg,B|A} = \bar{X}_{B|A} \beta_{B|A} \quad (12)$$

and

$$\beta_{B|A} = (X_{B|A}' W X_{B|A})^{-1} X_{B|A}' W Y . \quad (13)$$

Hence the bias estimator in (7) may be expressed alternatively as

$$\hat{B}(\bar{y}_{reg,A}) = -\bar{y}_{reg,B|A} = -\bar{X}_{B|A} \beta_{B|A} . \quad (14)$$

If β in (2) is partitioned as $\beta = (\beta_A' \; \beta_B')'$ then it follows (eg from a weighted version of Theorem 3.7(ii) of Seber, 1977) that $\beta_{B|A} = \beta_B$ and moreover if model (6) holds that the conditional bias of $\bar{y}_{reg,A}$ given the x_i is

$$E(\bar{y}_{reg,A} - \bar{y}_{reg} \mid x_i) = -\bar{X}_{B|A} \beta_B .$$

In so far as $\hat{\beta}$ is the preferred estimator of β we argue that the bias estimator in (7), which reduces to $-\bar{X}_{B|A} \hat{\beta}_B$, is the preferred estimator of the bias $-\bar{X}_{B|A} \beta_B$.

To obtain the variance estimator $v[\hat{B}(\bar{y}_{reg,A})]$ we proceed by linearising the regression estimators in the same way as in a derivation of the variance estimator in (5) to give

$$\bar{y}_{reg,A} - E(\bar{y}_{reg,A}) = n_r^{-1} \sum_r g_{Ai} e_{Ai} \quad (15)$$

$$\bar{y}_{reg} - E(\bar{y}_{reg}) = n_r^{-1} \sum_r g_i e_i , \quad (16)$$

where $e_{Ai} = y_i - x_{iA} E(\hat{\beta}_A)$ and $e_i = y_i - x_i E(\hat{\beta})$.

Subtracting (16) from (15) leads to

$$[\bar{y}_{reg,A} - \bar{y}_{reg}] - [E(\bar{y}_{reg,A}) - E(\bar{y}_{reg})] = n_r^{-1} \sum_r (g_{Ai} e_{Ai} - g_i e_i) \quad (17)$$

which suggests as the last term of (8) the following estimator

$$v[\hat{B}(\bar{y}_{reg,A})] = \sum_r (g_{Ai} \hat{e}_{Ai} - g_i \hat{e}_i)^2 / [n_r (n_r - J_B)] \quad (18)$$

where $\hat{e}_i = y_i - x_i \hat{\beta}$.

To summarize, the estimated MSE of $\bar{y}_{reg,A}$ is given in (8) with the terms on the right-hand side given by (5), (7) and (18).

3. Variable Selection Procedures

As a first variable selection approach, denoted BEST SUBSET, we propose to select the subset A of variables for which $\hat{MSE}(\bar{y}_{reg,A})$ is minimum. To apply the BEST SUBSET approach requires calculating (8) for all 2^{J-1} possible subsets of auxiliary variables (assuming a constant is always included), which may be very onerous computationally as J increases.

NS found that in the case of simple random sampling little precision was lost by using forward selection compared to a best subset approach. As our second variable selection method we therefore propose the following FORWARD SELECTION approach.

Begin with the subset A containing only the constant term ($x_{iA} = 1$) as an auxiliary variable and compute

$\hat{MSE}(\bar{y}_{reg,A})$. Now let B_k denote the kth variable in B to be considered for inclusion and set $C(k)=A \cup B_k$. Compute $\hat{MSE}(\bar{y}_{reg,C(k)})$ for every k and determine its minimum value, $\hat{MSE}(\bar{y}_{reg,C(K)})$ say, where B_K is the variable for which the estimated MSE is minimum. If $\hat{MSE}(\bar{y}_{reg,C(K)}) < \hat{MSE}(\bar{y}_{reg,A})$ then set A = C(K) and B = B - B_K, and proceed with another step of this algorithm. If $\hat{MSE}(\bar{y}_{reg,C(K)}) \geq \hat{MSE}(\bar{y}_{reg,A})$ then stop and use only the variables in A.

As our third approach, we consider simplifying the MSE estimator in (8). Note first that under the linear model in (6) with constant weights w_i and independent observations with constant error variance the least squares estimators β_A and $\beta_{B|A}$ are uncorrelated given the x_i. Under this approximating assumption we may therefore write, using (11),

$$\hat{MSE}(\bar{y}_{reg}) = v(\bar{y}_{reg,A}) + v[\hat{B}(\bar{y}_{reg,A})]$$

Subtracting from (8) gives

$$\hat{MSE}(\bar{y}_{reg,A}) - \hat{MSE}(\bar{y}_{reg}) = \hat{B}(\bar{y}_{reg,A})^2 - 2\,v[\hat{B}(\bar{y}_{reg,A})]$$

If again we follow a model-based approach conditional on the x_i we obtain from (14)

$$\hat{MSE}(\bar{y}_{reg,A}) - \hat{MSE}(\bar{y}_{reg}) = \bar{X}'_{B|A}\,(\beta_{B|A}\,\beta'_{B|A} - 2\,var(\beta_{B|A}))\,\bar{X}_{B|A}$$

which reduces to

$$\hat{MSE}(\bar{y}_{reg,A}) - \hat{MSE}(\bar{y}_{reg}) = \bar{X}^2_{B|A}\,(t^2_{B|A}-2)\,var(\beta_{B|A})$$

where

$$t_{B|A} = \beta_{B|A}/[var(\beta_{B|A})]^{1/2}$$

in the case when J_B=1. This suggests a very simple forward selection approach. Suppose as a very crude approximation that for any given subset A, adding any further auxiliary variable will remove any bias. Then the estimated MSE will fall by adding a given variable only if the associated $t^2_{B|A}$ is greater than 2. The quantity $t^2_{B|A}$ is simply the usual F-to-enter statistic commonly used in stepwise regression. Thus as a highly simplified (but easy to apply) approach we consider a STANDARD FORWARD SELECTION approach where the critical value of F is set to equal 2.

4. Simulation Study

In order to assess the performance of such variable selection procedures, a numerical simulation exercise was carried out based on a population consisting of 953 records for heads of household (HH) interviewed during the 1988 Test Population Census of Limeira, São Paulo state, Brazil. These records include all heads of household from enumeration areas 1 to 40 who filled in a long (sample) form.

The main purpose of this exercise is to assess how the choice of auxiliary variables affects the performance of regression estimators when they are used not only to improve precision but also to correct for nonresponse. The target survey variable y is total income, and the auxiliary variables available, denoted $x_{(1)}$ to $x_{(11)}$, are described in Table 1 in the Appendix.

A standard saturated linear regression model to predict y with $x_{(1)}$ to $x_{(11)}$ as explanatory variables was fitted using all the population records. The resulting fitted model displayed an adjusted R^2 of 0.8322, thus yielding reasonably high explanatory power. However, the fitted saturated model might be criticised as not parsimonius. A standard stepwise model search carried out considering all population records would lead to a "smaller" model, with only $x_{(11)}$, $x_{(1)}$ and $x_{(5)}$ included, having a slightly higher adjusted R^2 of 0.8329.

Despite this criticism, the saturated model was used here to generate predicted values for y. This decision was taken because a nonresponse mechanism that depends on all auxiliary variables available was

desired, such that the bias induced by nonresponse could in principle be removed by a saturated regression estimator, as hypothesized in the theoretical development of estimators for the bias of the regression estimator based on subsets of the auxiliary variables. The idea is to use the predicted values $\hat{y}$ based on the saturated regression model as input to a logit-type model for the probability that nonresponse occurs for each unit, given that it has been selected for the sample. The nonresponse probabilities were thus generated for all the population units by

$$1 - Pr(i \in r \mid i \in s) = \frac{\exp(a + bd_i)}{1.5[1 + \exp(a + bd_i)]} \quad (19)$$

where a=0, b=0.5 and the d_i are the standardised predicted values $\hat{y}_i$ based on the saturated model. Figure 1 (Appendix) displays the generated nonresponse probabilities for all the population units. The minimum and maximum are 0.2678 and 0.6648, respectively, with a median of 0.3083. This model assigns higher nonresponse probabilities for units with higher predicted income.

The simulation study proceeded by selecting 1,000 simple random samples of size 100 from the population, and then determining for each sample which units would be respondents. A sample unit would be selected as a respondent if a uniform[0,1] pseudo random number was greater than its corresponding nonresponse probability. Because this process was applied independently for each sample, a given unit might be a respondent for one sample and a nonrespondent for another. The reason for using this two stage process was that estimators which adjust for nonresponse using auxiliary information at the complete sample level might be of interest later, although for the current exercise this situation was not considered. Observed response rates for the 1,000 simple random samples of size 100 averaged 67%.

Five estimation strategies were applied for each sample replicate, all treating the basic weights w_i as constant for all sample units. The "sample mean of respondents" estimator $\bar{y}$ (denoted SM) and the saturated regression estimator $\bar{y}_{reg}$ (denoted SR) were computed as benchmarks. Three other strategies based on regression estimation after subset selection were also applied: the BEST SUBSET selection (denoted BS) and FORWARD SELECTION (denoted FS) procedures based on the MSE estimator given in (8), and a STANDARD FORWARD SELECTION (denoted SFS) procedure based on standard F-to-enter selection criteria, as implemented in SAS Proc Reg (see SAS, 1990). Table 2 in the Appendix displays the simulation estimates of the bias (absolute and relative) and mean squared error for each of these five estimation strategies.

The SM estimator displayed substantial bias in estimating the population mean $\bar{Y}$ (which is 150.4633). This was expected in view of the nonresponse mechanism imposed, for which units with higher predicted income $\hat{y}$ (and hence y) have higher nonresponse probabilities. The saturated regression estimator SR provided a substantial reduction in the MSE compared to the SM estimator. There is no evidence of a bias in the SR strategy which suggests that the assumption underlying the bias estimator in (7) is reasonable.

Regression estimation after BEST and FORWARD subset selection (BS and FS) performed very similarly and displayed negligible relative bias (-0.36%), while at the same time offering modest improvement in MSE over SR. Because FS is cheaper to compute, this method might be preferred over BS in applications, as already noted by NS. Finally, SFS also displayed no noticeable bias, with slightly higher MSE when compared to BS and FS, while still improving over SR.

Closer examination of the subsets selected by BS and FS revealed that the two approaches coincided in 69.6% of the samples, which helps explain their very similar performance. It is also worth noting that SFS generally selected smaller subsets, with 80% of samples yielding up to two variables included, whereas the other subset selection methods included an average of 6 variables in the model. This was expected because the selection criterion considered in this approach only takes variance into account, but not nonresponse bias, which is expected to be bigger for smaller subsets due to the nature of the nonresponse mechanism imposed.

Other important aspects of the proposed estimation procedures are variance estimation and coverage properties. NS noted that variance estimation may be difficult after variable selection since variance estimators may underestimate the MSE. Average of mean squared error estimates and empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each estimation strategy. The variance estimator used with SM, SR and SFS was the variance estimator given by (5), whereas

the minimised estimated MSE given by (8) was used with BS and FS. Results are given in Table 3 in the Appendix.

The underestimation of the MSE is severe for SM, particularly since the bias is substantial for this strategy. This is also reflected by the substantial undercoverage for this procedure. All the other procedures have similar coverage rates (around 85-86%) which are below the target 95% nominal level, as well as similar levels of underestimation (around 75-77%) of the corresponding simulation MSEs.

5. Conclusions

The results obtained here are similar to those found in NS for the simpler case when there is no nonresponse.They indicate that, in some instances, variable selection may be a useful strategy for improving precision over standard fixed subset regression estimation procedures, despite the higher associated computational costs. It is also worth noting that the incorporation of a bias component due to differential nonresponse into the MSE estimator used for variable selection is worth considering, at least in cases where the saturated regression estimator may be assumed to be approximately unbiased.

It is also worth noting that variance (MSE) estimation becomes more difficult after variable selection, although in the present case the saturated regression approach was not superior to regression estimation after variable selection according to this criterion.

Further research is still needed to develop better variance (MSE) estimators, as well as to assess the effect of using unequal weights w_i.

References

Bethlehem, J. G. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics, 4*, 251-260.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1983) Some uses of statistical models in connection with the nonresponse problem. In W.G. Madow and I. Olkin (eds) *Incomplete Data in Sample Surveys*, Vol 3, New York: Academic Press, 143-160.

Fuller, W.A., Loughlin, M.M. and Baker, H.D. (1994) Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology, 20*, 75-85.

Kalton, G.and Maligalig, D.S. (1991) A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*, US Bureau of the Census, Arlington VA, 409-428.

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

Nascimento Silva, P.L.D. and Skinner, C.J. (1997) Variable selection for regression estimation in finite populations. *Survey Methodology, 23*, 23-32.

Särndal, C.E. and Swensson, B. (1987) A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Int.Statist.Rev., 55*, 279-294.

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

SAS Institute Inc. (1990). SAS/STAT User's Guide (version 6, vol. 2, 4th ed.). Cary, NC: SAS Institute Inc.

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

Tremblay, V. (1986) Practical criteria for definition of weighting classes. *Survey Methodology, 12*, 85-97.

APPENDIX

Figure 1 - Generated nonresponse probabilities

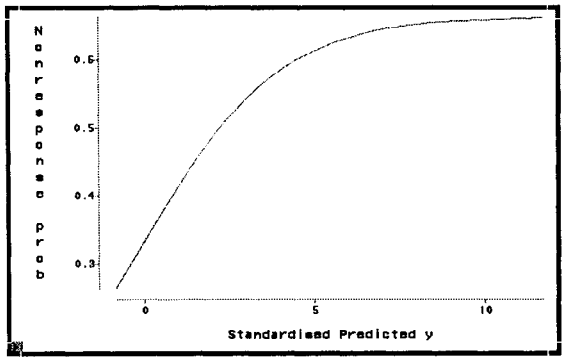

Table 1 - List of Auxiliary Variables Available

Variable Label	Variable Description
$x_{(1)}$	Indicator Sex = Male
$x_{(2)}$	Indicator Age <= 35
$x_{(3)}$	Indicator 35 < Age <= 55
$x_{(4)}$	Total Rooms in HH
$x_{(5)}$	Number of Bathrooms in HH
$x_{(6)}$	Indicator HH = Owned
$x_{(7)}$	Indicator HH = House
$x_{(8)}$	Indicator of Car in HH
$x_{(9)}$	Indicator of Colour TV in HH
$x_{(10)}$	Years of Study of Head of HH
$x_{(11)}$	Proxy Income

Table 2 - Bias, Relative Bias and Mean Squared Error estimates for alternative estimation strategies

Estimation Strategy	Bias	Relative Bias (%)	MSE
SM - Sample mean of Respondents	-17.88	-11.88	757.45
SR - Saturated Regression Est.	0.03	0.02	172.79
BS - Best Subset Regression Est.	-0.54	-0.36	143.27
FS - Forward Selection Regression Est.	-0.54	-0.36	143.50
SFS - Standard Forward Selection	0.36	0.24	152.32

Table 3 - Empirical coverage rates for nominal 95% confidence intervals and average of Mean Squared Error estimates for alternative estimation strategies

Estimation Strategy	Empirical Coverage Rate (%)	Average of MSE Estimates
SM - Sample mean of Respondents	72.1	429.05
SR - Saturated Regression Est.	86.1	130.30
BS - Best Subset Regression Est.	85.2	109.99
FS - Forward Selection Regression Est.	85.0	111.58
SFS - Standard Forward Selection	85.4	115.03