REPRESENTATIVE BOUNDED-LEVERAGE REGRESSION ESTIMATION FOR FINITE POPULATIONS

Johan C. Akkerboom and Nico J. Nieuwenbroek, Neth. Centr. Bur. of Stat.
Akkerboom, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands

KEY WORDS: Modified regression weights, design robustness

1. Introduction and summary

The use of auxiliary variables for weighted estimation of finite population totals or means is commonly advocated for reasons of bias and/or variance reduction. In practice, good predictors of the target variable(s) and/or the response mechanism may be hard to obtain. In such cases one often deals with categorical auxiliary variables only, and in any case the weighting procedure may amount to little more than standardization with respect to known population means. Standardization may have its price, however, in that the resulting weights vary too much. The variance of a weighted estimator may be unduly increased as compared with the corresponding unweighted estimator. Such an increase may be counteracted to a certain extent by truncating the weights, cf. Potter (1988,1990).

In the paper we start from the 'π-inverse weighted least squares estimator' $\hat{\bar{y}}_{LS}$ for the population mean and modify it into a general regression (GR-) estimator $\hat{\bar{y}}_{GR}$ with built-in truncation procedure, namely Potter's 'squared weight contribution to mean squared weight (NAEP) procedure'. The estimation weights v_i in the resulting *bounded leverage (BL-) estimator* $\hat{\bar{y}}_{BL}=\sum_i v_i y_i$ are bounded in absolute value, while $\hat{\bar{y}}_{BL}$ retains some general properties of $\hat{\bar{y}}_{LS}$, being 'in simple projection form', 'asymptotically design unbiased' (ADU), and representative with respect to the covariates.

In §2 we derive the BL-estimator as a 'x-outlier-robust' modification of a given GR-estimator. In §3 we describe the corresponding algorithm. In §4 we discuss related approaches in the survey literature.

2. Robust general regression estimators

Given a population of size N of (y_k, x_k)-pairs of target variable and covariate values, let (y_i, x_i), $i=1,\ldots,n$, be a sample of effective size n drawn from that population with inclusion probabilities π_k, $k=1,\ldots,N$ (x_k is 1-by-p). We use the notations $\Lambda=\mathrm{diag}(\lambda_i)$ with $\lambda_i=1/\pi_i$, $y=(y_1,\ldots,y_n)^T$, X for the n-by-p sample covariate matrix with i-th row x_i, $Y_N=(y_1,\ldots,y_N)^T$, X_N for the N-by-p design matrix with k-th row x_k, and 1_n and 1_N for vectors with all entries equal to 1. Our aim is to obtain a GR-estimator for the finite population total $Y=1_N^T Y_N=\sum_k y_k$ or the finite population mean $\bar{Y}=Y/N$ with good design properties, say with small design mean squared error. For the sake of obtaining a uniform set of estimation weights, regardless of y, we will not explicitly account for design bias. Instead, we aim to reduce the design variance by eliminating weights with large absolute value due to 'x-outliers'.

We will use the model

$$y_k = x_k\beta+\varepsilon_k \quad (k=1,\ldots,N), \quad E_\eta \varepsilon_k=0,$$

$$\mathrm{var}_\eta \varepsilon_k=\sigma^2 a_k, \quad E_\eta \varepsilon_j \varepsilon_k=0 \ (j\neq k), \qquad (1)$$

for the (y_k,x_k)-pairs, where the x_k and ε_k are assumed to be independently distributed, such that $x_k\beta$ is distributed around $\xi\beta=\mu$ and ε_k is distributed around 0 (a_k known, σ^2 unknown). In matrix notation we have $y_N=X_N\beta+\varepsilon_N$, with $E_\eta Y_N=X_N\beta$ and $E_\eta\varepsilon_N\varepsilon_N^T= =\sigma^2 V$, where $\varepsilon_N=(\varepsilon_1,\ldots,\varepsilon_N)^T$ and $V=\mathrm{diag}(a_k)$. The x_k are assumed to be known, or at least the p population means contained in the row vector $\xi=N^{-1}1_N^T X_N$. We focus on the case with intercept term, such that 1_N belongs to the column space of X_N ($x_{k1}=1$ for all k; X_N of full rank). If all covariates are categorical, then the resulting estimator will amount to (semi-)poststratification, cf. Bethlehem and Keller (1987).

The superpopulation model (1) serves to (a) lend support to taking the familiar class of general regression (GR-) estimators as a point of departure, and to (b) apply concepts like 'influence' and 'robustness' to such estimators, cf. Tam (1988) and Smith (1990). For any regression coefficient estimator $\hat{\beta}$, the GR-estimator for $\bar{Y}$ is given by

$$\hat{\bar{y}}_{GR} = \hat{\bar{y}}_{HT}+(\xi-\hat{\bar{x}}_{HT})\hat{\beta} =$$

$$= N^{-1}\{1_n^T\Lambda y+(1_N^T X_N-1_n^T\Lambda X)\}\hat{\beta}, \quad (2)$$

where $\hat{\bar{y}}_{HT}=N^{-1}\sum_i y_i/\pi_i$ is the Horvitz-Thompson estimator for $\bar{Y}$, and $\hat{\bar{x}}_{HT}$ the row of Horvitz-Thompson estimators for the ξ_j. In practice, the a_k will usually not be known to any reasonable accuracy, so that we simply put $a_k=1$ for all k ($V=I_N$).

Consider the π^{-1}-weighted least squares (LS-) estimator $\hat{\bar{y}}_{LS}$ with $\hat{\beta}=\hat{\beta}_{LS}=(X^T\Lambda X)^{-1}X^T\Lambda y$. One can write $\hat{\bar{y}}_{LS}=\sum_i w_i y_i$ where the estimation weights are given by

$$w_i = N^{-1}\lambda_i+\lambda_i(\xi-\hat{\bar{x}}_{HT})(X^T\Lambda X)^{-1}x_i^T =$$

$$= \lambda_i\xi(X^T\Lambda X)^{-1}x_i \quad (i=1,\ldots,n), \quad (3)$$

so that w_i is the sum of the 'sample weight' $1/(N\pi_i)$ and an 'adjustment weight', say α_i, associated with $\hat{\beta}_{LS}$. The second equality in (3) holds because $\sum_i \lambda_i e_i = 0$ for the sample residuals $e_i = y_i - x_i\hat{\beta}_{LS}$, or

$$1_n^T \Lambda(y - X\hat{\beta}_{LS}) = 0. \qquad (4)$$

This means that $\hat{\bar{y}}_{LS}$ can be written in *simple projection form* as

$$\hat{\bar{y}}_{LS} = \xi\hat{\beta}_{LS} = N^{-1}\sum_{k=1}^{N}\hat{y}_k, \qquad (5)$$

where $\hat{y}_k = x_k\hat{\beta}_{LS}$ is the ordinary linear model estimate of y_k. This form is convenient because the influence of observations for the i-th sample element on the outcome of $\hat{\bar{y}}_{LS}$ can be expressed in terms of their influence on $\hat{\beta}_{LS}$ only, cf. Smith (1990). Moreover, '$\sum_i \lambda_i e_i = 0$' offers protection against the simultaneous occurrence of a large residual e_i and a large sample weight $N^{-1}\lambda_i$.

Under certain regularity conditions, Särndal (1980) shows that with $\hat{\bar{y}}_{LS}$ we have a GR-estimator $T = T(y) = \sum_i w_i y_i$ with the property that

(i) T is *ADU*, that is

$\lim_{n,N\to\infty} E_p(T - \bar{Y}) = 0$.

Moreover, when c_j denotes the j-th column of X, we have that

(ii) T is *representative*, that is

$T(c_j) = \sum_i w_i x_{ij} = \xi_j$ for $j = 1,\ldots,p$,

and, as we already saw,

(iii) T is in *simple projection form*, that is $T = N^{-1}\sum_k x_k\hat{\beta}$.

Here i and ii are typical for 'model robustness' in the sense of Tam (1988). Property iii is not sufficient to guarantee that $\hat{\bar{y}}_{LS}$ is 'design-robust', in the sense that the (sample based) influence function $IF(y_i, x_i; \hat{\bar{y}}_{LS})$ is unbounded both in y_i and x_i, cf. Cook and Weisberg (1982). In fact we have

$$IF(y_i, x_i; \hat{\bar{y}}_{LS}) \propto \lambda_i \xi (X^T \Lambda X)^{-1} x_i^T e_i =$$

$$= w_i e_i \quad (i = 1,\ldots,n). \qquad (6)$$

This undesirable property relates to the fact that $\hat{\beta} = \hat{\beta}_{LS}$ is the solution to the normal equations $X^T \Lambda e = 0_p$, in which large residuals are weighted as heavily as small ones ($e_i = y_i - x_i\hat{\beta}$; $e = (e_1,\ldots,e_n)^T$).

Now we sketch how $\hat{\beta}_{LS}$ can be modified in such a way that the influence of the x_i is bounded, while properties i-iii are retained. Note that we refrain from accounting for 'y-outliers' by bounding the influence of residual. Focusing on covariate influence, let us turn to the class of so-called π^{-1}-weighted GR-estimators $\hat{\bar{y}}_{PI}$ of the form (2) with

$$\hat{\beta} = \hat{\beta}_{PI} = (Z^T X)^{-1} Z^T y, \quad \text{with}$$

$$\text{(a) } \operatorname{rank}(Z^T X) = p, \text{ and} \qquad (7)$$

$$\text{(b) } Zc = \Lambda 1_n \text{ for some p-vector c.}$$

Here Z is some n-by-p matrix that may depend on the π_i and x_i. Note that $\hat{\bar{y}}_{LS}$ is obtained for $Z = \Lambda X$. Now $\hat{\bar{y}}_{PI}$ is obviously representative, while it will be ADU under mild regularity conditions regarding Z.

As $\hat{\beta}=\hat{\beta}_{PI}$ is the solution to the normal equations $Z^T e=0_p$, condition b in (7) implies that $\hat{\bar{y}}_{PI}=\xi\hat{\beta}_{PI}$. We aim to choose Z in such a way that covariate influence is bounded without affecting conditions a and b in (7). In §3 we take $Z=(\Lambda 1_n|\Lambda U X_2)$, which is obtained by partitioning $X=(1_n|X_2)$ into 1_n and the n-by-(p-1) 'substantial covariate matrix' X_2. Here $U=\mathrm{diag}(\upsilon_i)$ is composed of iteratively computed correction factors υ_i. Equivalently, the p-fold system

$$\begin{cases} \sum_i \lambda_i e_i = 0, & (8a) \\ \sum \lambda_i \upsilon_i e_i x_{ij} = 0,\ j=2,\ldots,p, & (8b) \end{cases}$$

has to be iteratively solved for β ($e_i=y_i-x_i\beta$), where the υ_i depend on restrictions on the squares of the modified estimation weights given by

$$v_i = N^{-1}\lambda_i+(\xi-\hat{\bar{x}}_{HT})(Z^TX)^{-1}z_i^T = \xi(Z^TX)^{-1}z_i^T \quad (i=1,\ldots,n), \quad (9)$$

with $\hat{\bar{y}}_{PI}=\sum_i v_i y_i$. Simple projection form is thus preserved by bounding the influence of the p-1 substantial covariates without affecting that of the dummies x_{i1}.

3. Enforcing a bound for the squared weights

The following algorithm can be used to enforce an upper bound K^2 for the squared estimation weights v_i^2 so as to obtain the BL-estimator $\hat{\bar{y}}_{BL}$ given by (2) and (7), cf. (9). We let the bound K^2 depend on $\mathrm{Var}_\eta\hat{\bar{y}}_{LS}=\sigma^2\sum_i w_i^2$, the model variance of the LS-estimator, and on a preset proportionality factor C (σ^2 is cancelled out).

Our BL-algorithm incorporates the NAEP-procedure, the squared weights being bounded by an upper bound nK^2 proportional to $\mathrm{Var}_\eta\hat{\bar{y}}_{LS}/\sigma^2$. Obviously there remains some arbitrariness in the choice of the proportionality factor C, cf. Potter (1990). Here careful judgment is required, because the weights should not be so rigorously adjusted that bias is unduly increased. (C should be large enough to ensure convergence; the υ_i should approach 1 as $n,N\to\infty$.) Putting $\sigma^2=1$, the algorithm runs as follows:

(0)-Put $\upsilon_i=1$ and compute
   $w_i=\lambda_i\xi(X^T\Lambda X)^{-1}x_i^T$ $(i=1,\ldots,n)$.
   -Compute $\mathrm{Var}_\eta\hat{\bar{y}}_{LS}=\sum_i w_i^2$; choose C, cf. $\mathrm{DEFF}_W=n(\sum_i w_i^2)/(\sum_i w_i)^2$.
   -Put $K^2=Cn^{-1}\mathrm{Var}_\eta\hat{\bar{y}}_{LS}$.

(1)-Put $\tilde{X}=(\tilde{c}_1,c_2,\ldots,c_p)$, where $c_j=(x_{1j},\ldots,x_{nj})^T$ is the j-th column of X $(j=2,\ldots,p)$ and $\tilde{c}_1=(\tilde{x}_{11},\ldots,\tilde{x}_{n1})^T$ is such that $\tilde{x}_{i1}=1/\upsilon_i$ $(i=1,\ldots,n)$.
   -Compute $Z=\Lambda U\tilde{X}$ and $d_i=|v_i|=|\xi(Z^TX)^{-1}z_i^T|$ $(i=1,\ldots,n)$.
   -Replace υ_i by $\min(1,fK/d_i)\upsilon_i$ $(i=1,\ldots,n;\ K=(K^2)^{½};$ f is some constant for speeding up

convergence, say f=0.99).

(2) Repeat Step 1 until the $|\upsilon_i|$ and/or the d_i do not change substantially (say ≤1%), or until the number of iterations attains a certain maximum.

(3) Compute $v_i = \xi(Z^T X)^{-1} z_i^T$ $(i=1,\ldots,n)$.

This bounded-leverage algorithm is analogous to that of Dorsett (1989) in which the bound nK^2 is set for $x_i(Z^T X)^{-1} z_i^T$, the 'leverage' of x_i.

The above algorithm results in a (virtually) continuous nondecreasing transformation of the least-squares weights (small 1-f>0). The transformation is not smooth. Smoothness can be obtained, however, in the manner of Huang and Fuller (1978) by replacing $\min(1, fK/d_i)$, the correction factor in Step 1, by some smooth nondecreasing function of K/d_i that equals 1 for $K/d_i \geq 1+\delta$ (for some $\delta>0$) and fK/d_i for $K/d_i \leq 1$. It may be computationally advantageous to center the z_{ij} in the system $Z^T e = 0_p$, except for the dummy covariate $(j=2,\ldots,p)$. For the n-by-p matrix $\Gamma = (\gamma_{ij})$ with $\gamma_{i1} = 0$ and $\gamma_{ij} = \xi_j/\upsilon_i$, $j=2,\ldots,p$, one has $\Gamma^T U \Lambda e = 0_p$ because $1_n^T \Lambda e = 0$ holds throughout the algorithm. Hence at each iteration of Step 1, and in the final estimation formula of Step 3, $\tilde{X}$ may be replaced by $\tilde{X} - \Gamma$ in $Z = \Lambda U \tilde{X}$, that is z_{ij} may be replaced by $z_{ij} - \lambda_i \xi_j$ $(i=1,\ldots,n;\ j=2,\ldots,p)$.

4. Discussion

The paper gives a method of bounding the influence of substantial covariates on the outcome of the general regression estimator, without reference to any specific target variable. Potter (1988, 1990) makes a distinction between two categories of post-design procedures for limiting or reducing the number and size of extreme estimation weights. One category concerns the separate bounding of successive weight components that arise during the weighting process. The other one concerns inspection, truncation and compensation procedures after the composite weights have been computed. Bounded leverage (BL-) regression estimation represents a blend of these categories, in the sense that the final BL-weights are obtained in a single process, in which truncation is automatically compensated for.

Another feature of BL-regression estimation is that properties of model robustness (representativeness and asymptotic unbiasedness) are combined, to a certain extent, with 'design robustness', cf. Kish (1977). One would, of course, guarantee 'design robustness' more fully by somehow bounding the influence of residual as well as covariate influence, cf. Dorsett (1989) and Smith (1990).

Thus an increase in bias might be prevented automatically. It may also be necessary to 'robustify' $(Z^TX)^{-1}$ and thus avoid convergence problems.

Other methods of modifying regression weights have been proposed. Huang and Fuller (1978) give an algorithm similar to the BL-algorithm, such that the adjustment weight $\alpha_i = \lambda_i \upsilon_i (\xi - \hat{\bar{x}}_{HT})(X^T U \Lambda X)^{-1} x_i^T$ is bounded relative to the sample weight $N^{-1}\lambda_i$, for iteratively computed correction factors υ_i. The purpose of Huang's and Fuller's method is somewhat different from ours. Moreover, simple projection form is not preserved and hence influence considerations are less straightforward. An algorithm satisfying requirements of both nonnegativity and variance reduction would be fairly welcome.

References

Bethlehem, J.G. and W.J. Keller (1987). Linear Weighting of Sample Survey Data, *Journal of Official Statistics* 3, 141-154.

Cook, R.D. and S. Weisberg (1982). Residuals and Influence in Regression, John Wiley, New York.

Dorsett, D. (1989). Bounded-leverage weights for robust regression estimators, *Commun. Statist. - Theory Meth.* 18, 2785-2800.

Huang, E.T. and W.A. Fuller (1978). Nonnegative regression estimation for sample survey data, in Proceedings of the Section on Survey Research Methods, American Statistical Association, 300-303.

Kish (1977). Robustness in survey sampling, *Bulletin of the International Statistical Institute* 47 (3), 515-528.

Potter, F.J. (1988). Survey of Procedures to Control Extreme Sampling Weights, in Proceedings of the Section on Survey Research Methods, American Statistical Association, 453-458.

Potter, F.J. (1990). A Study of Procedures to Identify and Trim Extreme Sampling Weights, in Proceedings of the Section on Survey Research Methods, American Statistical Association, 225-230.

Särndal, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling, *Biometrika* 67 (3), 639-650.

Smith, J.W. (1990). The use of auxiliary data in robust univariate mean estimation, *Comm. Statist. - Simula.* 19, 787-807.

Tam, S.M. (1988). Some Results on Robust Estimation in Finite Population Sampling, *Journal of the American Statistical Association* 83, 242-248.