Division of Research
Graduate School of Business Administration
The University of Michigan

August 1, 1980

Robust Sampling Designs Using Several Auxiliary Variables

Working Paper No. 227

Roger L. Wright

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

ABSTRACT

Strategies are investigated for large-scale surveys of populations having known auxiliary variables related to the target variable through a linear superpopulation model. Strategies which combine WLS estimators with varying probability sampling designs are evaluated using criteria that integrate sampling and model-based considerations. The best robust strategy is found to incorporate a new WLS estimator and a design which generalizes Neyman allocation. This strategy is typically much more efficient than robust strategies using OLS or BLU estimators. The best robust strategy can be approximated by strategies using strongly stratified sampling.

KEY WORDS: Balanced sampling; Multiple regression models; Robustness; Stratification; Superpopulation models; Unequal probability sampling.

Author's Footnote:

Roger L. Wright is Associate Professor of Statistics, Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI 48109.
This paper was prepared with the support of the U.S. Department of Energy, Grant No. DE-FG02-80ER10125. However, any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not necessarily reflect the views of D.O.E.

The author wishes to thank K.R.W. Brewer, W.A. Ericson, and D.J. Aigner for valuable comments on an earlier version of this paper.

I. INTRODUCTION

Sampling strategies have conventionally combined sampling designs and estimators which minimize bias and provide reasonable efficiency with minimal assumptions about population distributions. In other fields, inference is more dependent on models, especially linear regression models. This paper identifies a class of strategies for using several auxiliary variables in WLS estimators which are asymptotically unbiased in a traditional sampling sense. By adopting a linear superpopulation model, a strategy can be constructed which is robust in the sense that it provides an asymptotically unbiased estimator regardless of the validity of the model, and is efficient in the sense that it minimizes the asymptotic variance if the model is accurate. This strategy is often substantially more efficient than robust strategies based on OLS or BLU estimators.

Following Royall (1976), consider a finite population comprised of N units labeled I=1,...,N. Unit I has known attributes given by the (k x 1) vector X_I in R^k, as well as an attribute Y_I which is the realization of a superpopulation random variable. For the purpose of design, assume the linear superpopulation model

$$Y_I = X_I'\beta + e_I \qquad (1.1)$$

$$\begin{aligned} E(e_I) = 0,\ E(e_I e_J) &= \sigma_I^2 > 0 \text{ if } J = I, \\ &= 0 \text{ otherwise,} \end{aligned}$$

with β unknown but σ_I known.

Two restrictions of (1.1) are of interest. Brewer (1979) has investigated the ratio model in which k=1 and $X_I > 0$. Our general analysis will prove to be most interesting in the case of the nonzero-intercept model in which the initial element of X_I is unity.

For any sample s, as an estimator, or a predictor under (1.1), of the total $Y = \sum^N_{I=1} Y_I$ we use

$$y^* = \sum_{I\in s} Y_I + \sum_{I\in \bar{s}} X_I'\hat{\beta} \qquad (1.2)$$

with

$$\hat{\beta} = \Big(\sum_{I\in s} W_I X_I X_I'\Big)^{-1} \sum_{I\in s} W_I X_I Y_I. \qquad (1.3)$$

Here it is assumed that the W_I and the sample s yield a nonsingular matrix $\sum_{I\in s} W_I X_I X_I'$; otherwise the W_I are to be chosen as part of the strategy.

Conventionally the W_I are chosen following criteria based on (1.1). The most widely recommended choice is $W_I = \sigma_I^{-2}$, so that y* is the best linear unbiased (BLU) predictor of Y conditional on (1.1) and s (Royall 1976). A second choice might be $W_I = 1$, giving the ordinary least squares (OLS) predictor y^*_{OLS}. A new choice of W_I will be proposed in Section 3. This strategy is generally superior to BLU or OLS when certain sampling considerations are included.

The precision of y* depends not only on the W_I but also on the sampling design. If (1.1) is known to be accurate, a sensible strategy is to choose the sample s to minimize the variance of y^*_{BLU}. For example, with the ratio model and $\sigma_I^2 = \sigma_o^2 X_I$, this strategy dictates that s be comprised of the n largest units in the population. In general, if the assumed model is inaccurate, y^*_{BLU} can be badly biased under this strategy.

This has stimulated interest in robust strategies that provide some degree of protection against model misspecification. Royall and Herson (1973a) and Scott, Brewer and Ho (1978) work with y^*_{BLU} for the ratio model and impose balance conditions on s that guarantee unbiasedness under a class of alternative models. Following their approach, y* is unbiased under the alternative model

$$Y_I = Z_I'\delta + u_I \text{ with } E(u_I) = 0$$

if s satisfies the balance condition

$$\sum_{I \in \bar{s}} X_I' \left(\sum_{I \in s} W_I X_I X_I' \right)^{-1} \sum_{I \in s} W_I X_I Z_I' = \sum_{I \in \bar{s}} Z_I'. \qquad (1.4)$$

The balanced sampling approach raises two questions:

1. How to select s satisfying (1.4) for one or more Z_I, and
2. How to offset the loss of efficiency in y* when s is forced to satisfy (1.4).

A partial answer is to use y^*_{BLU} with stratified sampling (Royall and Herson 1973b) or varying probability sampling (Scott et al. 1978).

A more complete answer leading in a new direction is given by Brewer (1979) for the ratio model. Instead of requiring model-unbiasedness under a specific class of alternative models, Brewer obtains robustness by imposing a condition which relates the W_I to the sampling design and which guarantees that y* is asymptotically design unbiased (ADU). Brewer then selects the sampling design to minimize the asymptotic variance of y*, say v(y*). (The asymptotic construction will be described in Section 2.) In particular, Brewer shows for the ratio model that:

1. y* is ADU if and only if the W_I are proportional to $(\pi_I^{-1} - 1)/X_I$ where π_I is the probability of selecting unit I.
2. If y* is ADU, then v(y*) is minimized with the sampling design $\pi_I = n\ \sigma_I/\sum^N{}_{J=1}\ \sigma_J$.
3. With this design,

$$v(y^*) = n^{-1}\ (\sum_{I=1}^{N}\ \sigma_I)^2 - \sum_{I=1}^{N}\ \sigma_I^2.$$

In the next two sections, Brewer's results will be generalized to the unrestricted multivariate model (1.1), and in particular, to the nonzero-intercept model.

2. ASYMPTOTICALLY DESIGN UNBIASED STRATEGIES

Following Brewer (1978), our asymptotic limits will be constructed as follows. For any positive integer m, consider m exact copies of the original finite population to form an aggregate population of mN units with total $Y_m = mY$. From each of these m populations, one sample is selected using fixed π_I; these m samples are considered together as a single aggregate sample. The estimator y_m^* is defined by applying (1.2) and (1.3) to the aggregate sample. To guarantee the existence of various limits, we assume that there exists a constant $\lambda > 0$ such that all characteristic roots of $\sum_{I \in s} W_I X_I X_I'$ are almost surely greater than λ.

With this construction, $\lim_{m\to\infty} E_p(y_m^*/m)$ exists and is equal to

$$\sum_{I=1}^{N} \pi_I Y_I + C' \sum_{I=1}^{N} \pi_I W_I X_I Y_I. \qquad (2.1)$$

Here E_p is the expectation based on the sampling design, and

$$C' = \sum_{I=1}^{N} (1-\pi_I) X_I' \left(\sum_{I=1}^{N} \pi_I W_I X_I X_I' \right)^{-1}. \qquad (2.2)$$

y* is said to be asymptotically design unbiased (ADU) if and only if $\lim_{m\to\infty} E_p(y_m^*/m)$ is equal to Y for any finite population. From (2.1), y* is ADU if and only if

$$\pi_I + \pi_I W_I C' X_I = 1, \qquad I=1,\ldots,N. \qquad (2.3)$$

Although (2.2-3) seem to provide a rather complex characterization of the W_I, suitable W_I can easily be constructed. Suppose D is any vector such that $D'X_I > 0$ for all I, and let

$$W_I = (\pi_I^{-1} - 1)(D'X_I)^{-1}. \qquad (2.4)$$

These W_I = satisfy (2.3) since (2.4) implies

$$C' = \sum_{I=1}^{N} \pi_I W_I D' X_I X_I' \left(\sum_{I=1}^{N} \pi_I W_I X_I X_I' \right)^{-1}$$

$$= D'.$$

This means that an ADU estimator y* can be constructed using (2.4) with any D for which $D'X_I > 0$ for all I.

Various choices of D generate simple classes of estimators in particular cases. While additional situations may arise and call for other choices of D, two cases are of interest here. For the ratio model, D is necessarily scalor so (2.4) implies that y* is ADU if and only if the W_I are proportional to $(\pi_I^{-1} - 1)/X_I$ as in Brewer (1979). Choosing $D = [\alpha^{-1}\ 0\ \ldots\ 0]'$, $\alpha > 0$, shows that y* is ADU for any model with a nonzero intercept if

$$W_I = \alpha(\pi_I^{-1} - 1), \quad \alpha > 0. \qquad (2.5)$$

3. BEST ADU STRATEGIES

Within the class of ADU strategies a useful performance measure is the asymptotic variance of y*, denoted v(y*). Here v(y*) is the asymptotic design-based expectation of the model-based mean squared error of y*. The asymptotic construction is the same as in Section 2. Specifically, v(y*) equals $\lim_{m\to\infty} EpE(y_m^* - Y_m)^2/m$.

v(y*) can easily be evaluated for any y* that is ADU. Using (1.1-3),

$$v(y^*) = \sum_{I=1}^{N} \pi_I W_I^2 \sigma_I^2 (C'X_I)^2 + \sum_{I=1}^{N} (1-\pi_I)\sigma_I^2 \qquad (3.1)$$

with C as in (2.2). If y* is ADU, then (2.3) can be used to simplify (3.1) giving

$$v(y^*) = \sum_{I=1}^{N} \pi_I^{-1}(1-\pi_I)^2 \sigma_I^2 + \sum_{I=1}^{N} (1-\pi_I)\sigma_I^2$$

$$= \sum_{I=1}^{N} (\pi_I^{-1} - 1)\sigma_I^2. \qquad (3.2)$$

The best ADU strategy is to choose the π_I to minimize (3.2). By Schwartz's inequality,

$$(\sum_{I=1}^{N} \sigma_I)^2 \leqslant \sum_{I=1}^{N} \pi_I \sum_{I=1}^{N} \sigma_I^2 \pi_I^{-1},$$

and $\sum_{I=1}^{N} \pi_I = n$, so

$$v(y^*) \geqslant n^{-1} (\sum_{I=1}^{N} \sigma_I)^2 - \sum_{I=1}^{N} \sigma_I^2. \qquad (3.3)$$

The lower bound (3.3) is achieved by a generalization of Neyman allocation,

$$\pi_I = n\sigma_I / \sum_{J=1}^{N} \sigma_J. \tag{3.4}$$

A best ADU strategy combines the sampling design (3.4) with any ADU estimator, denoted y^*_{BDU}, giving

$$v(y^*_{BDU}) = n^{-1}\left(\sum_{I=1}^{N} \sigma_I\right)^2 - \sum_{I=1}^{N} \sigma_I^2. \tag{3.5}$$

4. ADU STRATEGIES FOR MODELS WITH NONZERO INTERCEPT

Throughout this section, assume the model (1.1) with a nonzero intercept, and consider the class of ADU strategies satisfying (2.5). Within this context, the best ADU strategy can be compared to more conventional ADU strategies employing OLS or BLU estimators. These comparisons show that the best strategy can greatly outperform the conventional strategies.

Consider first an OLS strategy with $W_I = 1$ and sample size n_o giving the estimator $y*_{OLS}$. If (2.5) is used to provide robustness, then π_I is uniformly n_o/N as in simple random sampling, and (3.2) implies that

$$v(y*_{OLS}) = \sum_{I=1}^{N} (N/n_o - 1)\sigma_I^2$$

$$= (N^2/n_o)(1 - n_o/N) \left(\sum_{I=1}^{N} \sigma_I^2/N \right). \qquad (4.1)$$

By comparing (4.1) with (3.5), we find that the OLS strategy using a simple random sample of size n_o provides the same asymptotic variance as the best ADU strategy with sample size $n = (\text{eff})n_o$. Here the asymptotic efficiency (eff) of $y*_{OLS}$ is

$$\text{eff} = \left(\sum_{I=1}^{N} \sigma_I \right)^2 / \left(N \sum_{I=1}^{N} \sigma_I^2\right)$$

$$= (cv^2 + 1)^{-1}, \qquad (4.2)$$

where cv is the coefficient of variation of σ_I throughout the population.

There is some evidence suggesting that cv is likely to be well in excess of unity in populations of interest, implying that the asymptotic efficiency of $y*_{OLS}$ compared to $y*_{BDU}$ is likely to be substantially below 50%.

Some empirical work with real populations is underway. In any case, y^*_{OLS} is efficient only if (1.1) is homoscedastic.

A robust BLU strategy has about the same (in)efficiency as the robust OLS strategy. Consider the BLU strategy with $W_I = \sigma_I^{-2}$ and sample size n_0. Assume also that the π_I are uniformly small so that (2.5) gives π_I approximately equal to $n_0\sigma_I^2/\sum_{J=1}^{N} \sigma_J^2$. Then (3.2) shows that $v(y^*_{BLU})$ is approximately equal to (4.1). This means that, given small π_I, the asymptotic efficiency of y^*_{BLU} is about the same as the asymptotic efficiency of y^*_{OLS}, (4.2), and is less than 100% if (1.1) is heteroscedastic.

The inefficiency of y^*_{BLU} for a heteroscedastic model may be surprising. It results from the poor sampling design that is used to provide robustness, namely sampling with probability proportional to σ_I^2. The model-inefficiency of y^*_{BDU} is more than compensated by the design-efficiency of sampling with probability proportional to σ_I. Simply stated, the BLU strategy is likely to yield a sample containing too many units with large σ_I. It is interesting to note that the inefficient BLU sampling design is a pps design if, as is often assumed, σ_I^2 is proportional to size.

5. STRONGLY STRATIFIED ROBUST DESIGNS

It is sometimes advantageous to approximate the best ADU strategy using stratification. Suppose that

$$\{S_h: \quad h=1,\ldots,H\}$$

is any stratification of the population. Let cv_h be the coefficient of variation of σ_I within the N_h units of stratum h, so that

$$1 + cv_h^2 = N_h \sum_{I\in S_h} \sigma_I^2 \, (\sum_{I\in S_h} \sigma_I)^{-2}. \tag{5.1}$$

While Dalenius and Hodges (1959), Cochran (1961) and others have been primarily interested in designs with small H, we will examine designs that are strongly stratified in the sense that max $\{cv_h: \; h=1,\ldots,H\} = \varepsilon$ is small.

Suppose that y* is ADU, with

$$\pi_I = n_h/N_h, \; I\in S_h. \tag{5.2}$$

Using (3.2),

$$v(y^*) = \sum_{h=1}^{H} (N_h/n_h - 1) \sum_{I\in S_h} \sigma_I^2. \tag{5.3}$$

As in Neyman allocation, (5.3) is minimized given n by choosing n_h proportional to $(N_h \sum_{I\in S_h} \sigma_I^2)^{1/2}$. With this, (5.3) becomes

$$v(y^*) = n^{-1}[\sum_{h=1}^{H} (N_h \sum_{I\in S_h} \sigma_I^2)^{1/2}]^2 - \sum_{I=1}^{N} \sigma_I^2$$

$$= n^{-1}[\sum_{h=1}^{H} (1+cv_h^2)^{1/2} \sum_{I\in S_h} \sigma_I]^2 - \sum_{I=1}^{N} \sigma_I^2$$

$$\leqslant (1 + \varepsilon^2)n^{-1} \, (\sum_{I=1}^{N} \sigma_I)^2 - \sum_{I=1}^{N} \sigma_I^2. \tag{5.4}$$

An almost equally efficient design is obtained with the substantially more convenient allocation rule

$$n_h = n(\sum_{I \in S_h} \sigma_I) / \sum_{I=1}^{N} \sigma_I. \qquad (5.5)$$

In this case,

$$v(y^*) = n^{-1} (\sum_{I=1}^{N} \sigma_I) \sum_{h=1}^{H} (1 + cv_h{}^2)(\sum_{I \in S_h} \sigma_I) - \sum_{I=1}^{N} \sigma_I{}^2$$

which is also bounded above by the right hand side of (5.4).

The factor $1 + \varepsilon^2$ in (5.4) limits the loss in efficiency in y* that comes from using (5.5) rather than (3.4). Since (5.4) depends on the stratified design only through ε, all strongly stratified designs are almost equally efficient given Neyman allocation or (5.5). So the choice of stratification is largely inconsequential, but a convenient criterion is to choose strata to equalize $\sum_{I \in S_h} \sigma_I$, so that (5.5) gives $n_h = n/H$. In the large scale surveys of interest, H can be chosen large enough so that ε is negligible.

While the best ADU strategy has been justified on asymptotic grounds, it may be that convergence to these asymptotic limits is accelerated by using strongly stratified designs. If this is true, then strongly stratified designs may perform well with moderate sample sizes.

REFERENCES

Brewer, K.R.W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," Journal of the American Statistical Association, 74, 911-915.

Cochran, W.G. (1961), "Comparison of Methods for Determining Stratum Boundaries," Bulletin of the International Statistical Institute, 38, 345-358.

Dalenius, T. and J.L. Hodges, Jr. (1959), "Minimum Variance Stratification," Journal of the American Statistical Association, 54, 88-101.

Royall, Richard M. (1976), "The Linear Least-Squares Prediction Approach to Two-Stage Sampling," Journal of the American Statistical Association, 71, 657-664.

Royall, Richard M. and Jay Herson. (1973a), "Robust Estimation in Finite Populations, I," Journal of the American Statistical Association, 68, 880-889.

______ (1973b), "Robust Estimation in Finite Populations, II: Stratification on a Size Variable," Journal of the American Statistical Association, 68, 890-893.

Scott, A.J., K.R.W. Brewer, and E.W.H. Ho. (1978), "Finite Population Sampling and Robust Estimation," Journal of the American Statistical Association, 73, 359-361.