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

February, 1981

Sample Design with Multivariate Auxiliary Information

Working Paper No. 247

Roger L. Wright

The University of Michigan

ABSTRACT

Strategies are investigated for planning large administrative sample surveys of populations having known auxiliary variables related to the target variable through a linear superpopulation model. Both model-based linear prediction strategies and design-based generalized regression strategies are imbedded within a class of strategies combining weighted least squares regression estimators and varying probability sample designs. Strategies are identified which provide asymptotically design unbiased (ADU) estimators regardless of the validity of the assumed model. The model-based asymptotic efficiency of these ADU strategies is related to the sample design. Practical stratified sampling plans are proposed which utilize inclusion probabilities related to a simple measure of the relevance of units. These plans generalize equal aggregate size rules for constructing stratified sampling plans. This methodology is illustrated in the context of utility load research and cost accounting.

KEY WORDS: Balanced sampling, Cost accounting, Load research, Regression estimators, 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 the D.O.E.

The author wishes to thank K.R.W. Brewer, Graham Kalton, and Leslie Kish for valuable comments on earlier versions of this paper.

1. INTRODUCTION

Often in management a project is undertaken to collect data on a sampling basis to augment an existing administrative database. For example, an accountant may need to estimate the current value of assets, a utility cost-of-service study may require estimates of the usage of electricity by time-of-day, or a marketing study may be undertaken to estimate the potential sales of a new product to established customers. Typically, additional relevant data, e.g. past sales of related products, are available within the administrative database for each unit in the population. This auxiliary information can be exploited to produce more reliable sample estimates. One approach is to use the sample data to estimate a regression model relating the target variable to the relevant auxiliary information, and then use the estimated regression equation to extend the target variable to the unsampled part of the population.

This use of auxiliary information can usually be anticipated when the project is planned. Past experience often indicates the character of the regression relationship to be expected. This expected regression relationship can be utilized to choose the size of the sample and to develop an efficient sampling procedure. In many cases, a single-stage sample can be selected directly from the frame provided by the administrative database, so that the sampling plan is almost completely characterized by the probability of including each unit in the sample. These inclusion probabilities can be effectively chosen in accordance with the relevance of each unit as determined by the expected regression model and the population characteristic to be estimated.

Planning these administrative sampling projects is similar in many respects to planning the establishment surveys of public agencies, e.g. the U.S. Bureau of Labor Statistics' Current Employment Survey. The differences are a

matter of degree--most importantly the greater reliance on the administrative database both as a sampling frame and as a basis for estimation through modeling. Moreover, while most of the public establishment surveys utilize rotating sample designs, these more complex designs are fairly rare in management applications.

The management sampling applications that are of interest, then, share the following characteristics:

...A single-stage, stratified sampling plan is to be used to select a sample s from a finite population comprised of N units labeled I = 1,...,N.

...The purpose of the project is to estimate a finite population characteristic of the form $\sum_{I=1}^{N} a_I y_I$ with a_I known. This may be the population total $\sum_{I=1}^{N} y_I$, a subclass total or mean, a difference between subclass totals or means, or even a more complex characteristic such as a finite population regression coefficient.

...Past experience suggests that the target variable y_I is closely related to a vector $X_I = (x_1 \ldots x_k)' \varepsilon R^k$ of k auxiliary variables which are known throughout the population. The relationship between the target variable y and the auxiliary variables is thought to be well described by a regression model ξ: $y_I = X_I'\beta + u_I$ with unknown $\beta \varepsilon R^k$. Under ξ, the u_I are assumed to be random variables satisfying $E_\xi(u_I) = 0$, $E_\xi(u_I^2) = \sigma_I^2 > 0$, and $E_\xi(u_I u_J) = 0$ if $I \neq J$.

...The σ_I can be regarded as known, at least up to a constant of proportionality. In practice the σ_I are often assumed to be proportional to some known measure of size. More generally, past experience may suggest a particular functional relationship between σ_I and multivariate auxiliary information, (Harvey, 1976).

...A sampling strategy is to be developed which is to be efficient in the sense that it can be expected to provide a highly reliable estimator of $\sum_{I=1}^{N} a_I y_I$ if ξ is accurate, but which is also robust in the sense that the estimator is not badly biased even if ξ is misspecified.

An example may be helpful. Under the Public Utility Regulatory Policies Act of 1978 (PURPA), U.S. electric power companies are required to estimate the total power usage of various classes of customers during certain peak hours. These estimates are used in cost-of-service studies, to allocate the cost of maintaining generation and transmission capacity to various customer classes.

Because hourly usage is not normally metered, peak usage is estimated on a sampling basis. Available sample data generally show a strong relationship between peak usage and monthly consumption, which is usually metered for the entire population. Other potential predictors of peak usage include annual consumption, local weather characteristics, and perhaps the price of electricity. The latter is especially relevant in rate experiments. In addition household income, composition, and appliance stock may be available on a double sampling basis. Some references are Aigner (1979), Aigner and Housman (1980), and Taylor (1977).

In these studies the marginal cost of each sample unit is several hundred dollars or more per year so there is ample motivation to make efficient use of relevant and available auxiliary information. However, since these estimates have a substantial impact on electricity prices, they must have strong credibility with the public. This credibility seems to be related to the robustness of the sampling strategy.

In applications like these, the conventional approach is to use stratification to bring auxiliary information into the estimator and possibly to

introduce varying inclusion probabilities in the sampling plan. A common practice is to establish a stratification on one or more auxiliary variables using the Dalenius and Hodges (1959) cumulative square root rule on each variable, and then to use Neyman allocation based on within-strata variances estimated from available sample data. Often, because of limited sample data, allocation is based on the within-strata variance of an auxiliary variable that is thought to be highly correlated with the target variable. Various aspects of this procedure have been discussed by Anderson, Kish and Cornell (1980), Cochran (1961), Rao (1977), and Singh (1971, 1975).

This procedure requires considerable care and judgment. In deciding on the number of auxiliary variables to be used in the stratification and the number of cutpoints for each such variable, assumptions must be made about the joint distribution of the target variable and the auxiliary variables. Within-strata variances are usually difficult to estimate, and the use of an auxiliary variable as a proxy for the target variable conflicts with the use of the same auxiliary variable as a stratification variable. Moreover, there is scarce justification for using the Dalenius-Hodges rule for multivariate stratification, but other alternatives are extremely cumbersome.

In the present paper, the auxiliary information is used in a multivariate regression estimator. Multiple regression provides a familiar, easily used, and extremely flexible tool for bringing auxiliary information into the analysis. Dummy variables can be used in the model to represent categorical information; this is analagous to analysis of variance models and generalizes the technique of stratification in sampling. Other suitable variables can be included in linear or perhaps quadratic form, as in analysis of covariance. Interaction variables offer additional flexibility that generalizes the distinction in the sampling literature between separate and combined estimators in stratification.

In part because of the great flexibility that is available, care must be taken in developing a suitable regression model, (Konijn, 1973, p. 131). Attention must be given to variable selection, multicollinearity, and identifying outliers. However, techniques for handling these problems are fairly well developed and are familiar to many analysts, (e.g. Belsley, Kuh and Welsch, 1980 and Hocking, 1976).

An impediment to wide use of regression estimators in sampling applications has been the apparent contradictions between model-based procedures and sampling considerations. To help reconcile several approaches, Section 2 formulates a class of multivariate regression estimators which includes the linear predictors of Royall (1970, 1971, 1976) and the generalized regression estimators of Cassel, Sarndal and Wretman (1976, 1977). Sampling strategies that integrate the choice of sampling plan and estimator are considered.

Emphasizing design-based considerations, Section 3 proposes that robustness be achieved by restricting the strategies to those that are asymptotically design unbiased (ADU), regardless of the model's validity. The subclass of ADU strategies is shown to be determined by an algebraic condition which is used in subsequent analysis and facilitates construction of specific ADU strategies.

In Section 4 the model is used to examine the asymptotic efficiency of ADU strategies. When the available auxiliary information is used in an ADU regression estimator, the main role of the sampling plan is to provide suitable inclusion probabilities. The optimal inclusion probabilities are determined by the heteroscedasticity in the model and the characteristic to be estimated, and are in fact proportional to $|a_I|\sigma_I$, called the relevance of I. A useful basis for evaluating varying probability sampling plans is the efficiency of an equal probability plan, e.g. a simple random or proportionately allocated sampling plan.

Section 5 shows that the ideal inclusion probabilities can be well approximated with a stratified sampling plan using a simple stratification based on relevance. The choice of strata boundaries is shown to be much less critical than in the conventional approach since only the residual variation is of concern, not the within-strata variation of the auxiliary variables. A simple rule is proposed for constructing administratively convenient strata, which generalizes the equal aggregate size recommendation of Hansen, Hurwitz and Madow (1953, pp. 215-219).

Section 6 provides two numerical illustrations drawn from utility rate research and cost accounting.

2. SAMPLING STRATEGIES

As suggested in the previous section, the finite population characteristic $\sum_{I=1}^{N} a_I y_I$ is to be estimated using observed y_I, $I \in s$, together with auxiliary information $X_I \in R^k$ known throughout the population. The basis for planning is the superpopulation regression model ξ: $y_I = X_I'\beta + u_I$ with $E_\xi(u_I) = E_\xi(u_I u_J) = 0$, $I \neq J$. Here $\beta \in R^k$ is unknown but the $\sigma_I^2 = E_\xi(u_I^2)$ are regarded as known. To estimate β, we use the class of weighted least squares estimators with weights $q_I > 0$: $\hat{\beta} = (\sum_{I\in s} q_I X_I X_I')^{-1} \sum_{I\in s} q_I X_I y_I$.

While an obvious estimator for y_I is $X_I'\hat{\beta}$, additional useful information may be extracted from the sample residuals. For all I, define $\hat{u}_I$ to be $y_I - X_I'\hat{\beta}$. For $I \in s$, $\hat{u}_I$ is the observed sample residual which is usually regarded as containing information about the accuracy of ξ. By defining additional weights $r_I > 0$ associated with the $\hat{u}_I$ we obtain the class of estimators $\sum_{I=1}^{N} a_I \hat{y}_I$ with $\hat{y}_I = X_I'\hat{\beta} + r_I \delta_I \hat{u}_I$. Here δ_I identifies the sample, i.e. $\delta_I = 1$ if $I \in s$, 0 otherwise. We regard a <u>sampling strategy</u> to be determined by a sampling plan characterized by the inclusion probabilities

$\pi_I = Pr(I \in s)$ together with an estimator determined by the choice of q_I and r_I, I=1,...,N.

Additional vector notation is useful. Define $a = [a_1 \ \cdots \ a_N]'$, $y = [y_1 \ \cdots \ y_N]'$, $u = [u_I \ \cdots \ u_N]'$, and $e = [1 \ \cdots \ 1]'$, all in R^N. Let $X = [X_1 \ \cdots \ X_N]'$, the (Nxk) matrix of auxiliary information. Also define the following (NxN) diagonal matrices: $\Sigma = diag(\sigma_I^2)$, $\Pi = diag(\pi_I)$, $Q = diag(q_I)$, $R = diag(r_I)$, and $\Delta = diag(\delta_I)$. In this notation, the model ξ is

$$y = X\beta + u, \ E_\xi(u) = 0, \ E_\xi(uu') = \Sigma, \text{ known.} \qquad (2.1)$$

A sampling strategy is characterized by the triplet (Π,Q,R). To estimate the population characteristic a'y, we use the estimator $a'\hat{y}$ with

$$\hat{y} = X\hat{\beta} + R\Delta\hat{u}, \qquad (2.2)$$

$$\hat{\beta} = (X'Q\Delta X)^{-1} X'Q\Delta y, \text{ and}$$

$$\hat{u} = y - X\hat{\beta}.$$

It is assumed that the sample size $n = \sum_{I=1}^{N} \pi_I$ is fixed, and that $X'Q\Delta X$ is nonsingular for all s with nonzero probability of occurrence.

Various subclasses of strategies (Π,Q,R) have been considered previously. Strategies with R = I, the identity matrix, will be called <u>linear prediction strategies</u>, (Royall, 1970, 1976; Scott and Smith, 1969; and Smith, 1976). The class of strategies with $\Pi > 0$ and $R = \Pi^{-1}$ will be called <u>generalized regression strategies</u> following Cassel, Sarndal and Wretman (1976), and Sarndal (1980). Strategies with R = 0 will be called <u>simple projection strategies</u>. Strategies can be further classified by Q. Important cases are the <u>BLU</u> strategies with $Q = \Sigma^{-1}$ and the <u>III</u> (Π-inverse) strategies with $\Pi > 0$ and $Q = \Pi^{-1}$.

As Holt and Smith (1979) note, preference for strategies depends upon a tradeoff between considerations derived from the model (2.1) and considerations derived from the sample design, Π. Under ξ, $a'y$ is regarded as a random variable to be predicted by $a'\hat{y}$. Conditional on the sample s, the best linear unbiased predictor is

$$\sum_{I\epsilon s} a_I y_I + \sum_{I\notin s} a_I X_I'\hat{\beta}$$

$$= a'\Delta y + a'(I-\Delta)X\hat{\beta}$$

$$= a'X\hat{\beta} + a'\Delta\hat{u}.$$

as determined by the BLU linear prediction strategy (Π,Σ^{-1},I), (Royall, 1976; Smith, 1976). Here the observed sample residuals are used for the sample cases but provide no information about the unobserved residuals. The sample design Π plays no role in this estimator.

The heavy reliance of BLU strategies on ξ bothers many samplers, (e.g. Hansen, Madow and Tepping, 1978). They seem to prefer ΠI strategies because $X'\Delta\Pi^{-1}X$ and $X'\Delta\Pi^{-1}y$ are design unbiased estimators of $X'X$ and $X'y$, (e.g Fuller, 1975; Jonrup and Rennermalm, 1976; Kish and Frankel, 1974; and Konijn, 1962).

Looking for another compromise, Brewer (1979) suggests a modified ΠI linear prediction strategy which uses $q_I = (\pi_I^{-1} - 1)/x_I$ and $r_I = 1$ for the ratio model (2.1), with $k = 1$ and $x_I > 0$.

A more fundamentally sampling-based position is taken by Cassel, Sarndal and Wretman (1976, 1977). They recommend the generalized regression estimator

$$\sum_{I\epsilon s} a_I\pi_I^{-1}y_I + (\sum_{I=1}^{N} a_I X_I' - \sum_{I\epsilon s} a_I\pi_I^{-1}X_I')\hat{\beta}$$

$$= a'\Pi^{-1}\Delta y + a'(X-\Pi^{-1}\Delta X)\hat{\beta}$$

$$= a'X\hat{\beta} + a'\Pi^{-1}\Delta\hat{u}.$$

Here the sample residuals are thought to be informative about the unobserved

residuals. As in the Hurwitz-Thompson estimator, the sample residuals are extended to the population by using the sampling design. With this approach, the choice of $\hat{\beta}$ seems to be less critical than with other estimators, and both BLU and III estimators have been suggested (Sarndal, 1980), as well as generalized ratios (e.g. Raj, 1965).

There remains considerable confusion about effective regression strategies, and this confusion has undoubtedly deterred their use. The choice of strategy seems to be unavoidably dependent upon a subjective evaluation of the credibility of the model, the character of the application, and the nature of the available data. A universally optimal strategy cannot be prescribed, but perhaps some of the issues can be clarified.

A substantial advantage of model-based analysis is the strong links that are established with linear statistical inference, (e.g. Rao, 1973). With the added definitions

$$\hat{C} = X(X'Q\Delta X)^{-1}X'Q, \text{ and} \qquad (2.3)$$

$$\hat{T} = \hat{C} + R - R\Delta\hat{C},$$

we have $X\hat{\beta} = \hat{C}\Delta y$, $\hat{y} = \hat{T}\Delta y$, $\hat{C}\Delta X = X$, and

$$I - \hat{T}\Delta = (I-R\Delta)(I-\hat{C}\Delta)$$

$$= (I-\Delta) + (I-\hat{T})\Delta.$$

The prediction error $a'y - a'\hat{y} = a'(I-\hat{T}\Delta)y$ reduces to $a'(I-\hat{T}\Delta)u$ under ξ since $(I-\hat{C}\Delta)X = 0$. This implies that $a'\hat{y}$ is a ξ-unbiased predictor of $a'y$, with the mean squared error

$$a'(I-\hat{T}\Delta)\Sigma(I-\hat{T}'\Delta)a$$

$$= a'(I-\Delta)\Sigma(I-\Delta)a + a'(I-\hat{T})\Sigma\Delta(I-\hat{T}')a.$$

Under any linear prediction strategy, $\hat{T}-I = (I-\Delta)\hat{C}$ so the mean squared error simplifies to

$$a'(I-\Delta)(\Sigma + \hat{C}\Sigma\Delta\hat{C}')(I-\Delta)a.$$

As mentioned, this is minimized by using a BLU linear prediction strategy, (Π, Σ^{-1}, I). Of course this is conditional on the sample s.

If (2.1) is believed to be accurate, the choice of strategy seems to reduce to the choice of sampling plan Π. For instance, working with the ratio model, Royall (1970) has shown that rather weak conditions on Σ imply that the mean squared error is minimized by systematically selecting the n largest units in the population.

Despite the optimal properties of such a strategy, many survey samplers find it unacceptable, (e.g. Hansen, Madow and Tepping, 1978). For example, few consumer advocates would accept a utility cost-of-service study which determines prices for electricity from the consumption patterns of the largest users in various classes.

The problem with optimal model-based strategies does not seem to be the use of (2.1). Godambe's work implies that suitable strategies can only be identified by utilizing some sort of assumptions about the population (Smith, 1976, p. 187). One approach is to utilize information incorporated in a prior distribution, (Ericson, 1969; Scott and Smith, 1969). A closely related approach that has practical appeal is to use the model ξ to provide this information, (Anderson, Kish and Cornell, 1980; Brewer, 1963; and Rao, 1970).

The real problem with optimal model-based strategies seems to be their potential bias if the assumed model is even moderately inaccurate. This concern has stimulated interest in robust strategies that provide some degree of protection against model misspecification. Royall and Herson (1973a,b) and Scott, Brewer, and Ho (1978) provide unbiasedness under a specified class of alternative models by imposing balance conditions on the sample s. Although these writers restrict themselves to the strategies (Π, Σ^{-1}, I) for the ratio model, it is easily seen that given any strategy (Π, Q, R), $a'\hat{y}$ is

unbiased under the alternative model $y = Z\gamma + v$ with $E(v) = 0$ if and only if s satisfies the balance conditions $a'(I-\hat{T}\Delta)Z = 0$, or equivalently

$$a'(\hat{T}-I)\Delta Z = a'(I-\Delta)Z.$$

For a linear prediction strategy, the balance condition simplifies to

$$a'(I-\Delta)\hat{C}\Delta Z = a'(I-\Delta)Z.$$

The balanced sampling approach raises three questions:

1. How to choose the relevant Z,
2. How to identify the set of samples that satisfy the balance conditions for Z, and
3. How to isolate the most suitable sample within the set of balanced samples.

The literature that addresses these issues shows an evolution toward a design-based viewpoint, although the model-based strategies (Π,Σ^{-1},I) are generally retained.

Just as advocates of model-based strategies have been led to recognize design considerations, survey samplers more comfortable with design-based inference acknowledge the potential importance of model-based planning, although they tend to stay with the design-based strategies (Π,Π^{-1},Π^{-1}) or $(\Pi,\Pi^{-1},0)$.

Brewer (1979) and Sarndal (1980) have begun a systematic reconciliation of these approaches. This paper attempts to extend and unify their work by studying the general class of strategies (Π,Q,R) in a fashion that integrates model-based and design-based considerations. The concept of asymptotic design unbiasedness is used in the place of balance to provide robustness, while the asymptotic model-based mean squared error is used to analyze efficiency. This analysis is carried out in Sections 3 and 4 using the convenient context of varying probability sampling, but Section 5 shows how these strategies can be implemented using conventional stratified sampling.

3. ASYMPTOTICALLY DESIGN-UNBIASED STRATEGIES

Samplers find great comfort, and appropriately so, in strategies that yield estimators that are design-unbiased regardless of the population, e.g. $\bar{y}$ calculated from a simple random sample. A bit more reluctantly, they have recognized the usefulness of an estimator such as a ratio estimator that may be biased but is asymptotically design unbiased (ADU). The concept of balance has been introduced in an attempt to meet these concerns and still retain most of the advantages of model-based planning and inference. A more direct approach advanced by Brewer (1979) is to examine model-based strategies that give ADU estimators regardless of the validity of the model. This seems to side-step the problems with balanced sampling and to meet the needs of samplers.

In dealing with finite population sampling, care must be exercised in defining the context of asymptotic analysis. For our purposes, it is inadequate to simply let n increase to N. Instead we let the population size and sample size both increase with the sampling fraction fixed.

To preserve the character of the original finite population, the population size is conceptually increased by considering an aggregate population of mN units comprised of m copies of the original population. These m copies are assumed to be identical with respect to the known auxiliary information X. For model-based analysis, (2.1) is used to generate m independent realizations of y, say y_j, $j = 1,\ldots,m$. However to make this section's analysis independent of the model, in this section the y_j are considered to be identical copies of the original y.

Under any strategy (Π,Q,R), an aggregate sample of mn units is selected from the aggregate population by selecting an independent sample s_j from each of the m copies of the population. For each s_j we construct an (NxN) indicator matrix $\Delta_j = \mathrm{diag}[\delta_I(s_j)]$.

An estimator of the aggregate population characteristic $a'y_m = m^{-1} \sum_{j=1}^{m} a'y_j$ is formulated by applying the chosen strategy to the aggregate sample. The estimator is defined to be $a'\hat{y}_m$ where

$$\hat{y}_m = m^{-1} \sum_{j=1}^{m} \hat{y}_j, \tag{3.1}$$

$$\hat{y}_j = X\hat{\beta}_m + R\Delta_j \hat{u}_j,$$

$$\hat{\beta}_m = \left(\sum_{j=1}^{m} X'Q\Delta_j X \right)^{-1} \sum_{j=1}^{m} X'Q\Delta_j y_j, \text{ and}$$

$$\hat{u}_j = y_j - X\hat{\beta}_m.$$

We also define

$$\hat{\Pi}_m = m^{-1} \sum_{j=1}^{m} \Delta_j, \tag{3.2}$$

$$\hat{C}_m = X(X'Q\hat{\Pi}_m X)^{-1}X'Q, \text{ and}$$

$$\hat{T}_m = \hat{C}_m + R - R\hat{\Pi}_m \hat{C}_m,$$

so that

$$\hat{y}_m = m^{-1} \hat{T}_m \sum_{j=1}^{m} \Delta_j y_j.$$

Using the assumption of this section that $y_j = y$, the population characteristic of interest is $a'y$ and $\hat{y}_m$ becomes $\hat{T}_m \hat{\Pi}_m y$. Moreover the assumption that $X'Q\Delta X$ is nonsingular for all samples with non-zero probability of occurrence implies that $\hat{C}_m$ is bounded. Using this, the strong law of large numbers, and the Helly-Bray Theorem (Rao, 1963, p. 117), we have

$$\lim_{m\to\infty} E_p(\hat{y}_m) \tag{3.3}$$

$$= \lim_{m\to\infty} E_p(\hat{T}_m \hat{\Pi}_m y)$$

$$= T\Pi y$$

with

$$C = \lim_{m\to\infty} E_p(\hat{C}_m)$$

$$= X(X'Q\Pi X)^{-1}X'Q, \text{ and}$$

$$T = \lim_{m\to\infty} E_p(\hat{T}_m)$$

$$= C + R - R\Pi C.$$

Here E_p represents expectation with respect to the sampling distributions determined by Π.

This motivates

Definition 1. The strategy (Π,Q,R) is asymptotically design unbiased (ADU) for the characteristic a if and only if $a'(I-T\Pi)y = 0$ for all $y \in R^N$.

An immediate consequence of this definition is that for any strategy (Π,Q,R) that is ADU for a, $\pi_I = 0$ implies $a_I = 0$. Any unit with both $\pi_I = 0$ and $a_I = 0$ is clearly irrelevant and can be eliminated from the population. Because we are primarily interested in ADU strategies, it is assumed henceforth that $\Pi > 0$.

An equivalent characterization of ADU strategies can be developed from the identity $I-T\Pi = (I-R\Pi)(I-C\Pi)$. Suppose initially that $Q > 0$ so that $Q\Pi$ defines an inner product over R^N. In this case $C\Pi = X(X'Q\Pi X)^{-1}X'Q\Pi$ is the orthogonal projector onto the linear manifold M(X) spanned by the column vectors of X, and $I-C\Pi$ is the projector onto the linear manifold orthogonal to M(X) with respect to the inner product $Q\Pi$, (Rao, 1973, p. 47).

Since

$$a'(I-T\Pi)y = a'(I-R\Pi)(Q\Pi)^{-1}Q\Pi(I-C\Pi)y,$$

(Π,Q,R) is ADU for a if and only if $(Q\Pi)^{-1}(I-R\Pi)a \in M(X)$, or equivalently, $(I-R\Pi)a = Q\Pi x$ for some $x \in M(X)$. The restriction $Q > 0$ can easily be relaxed, giving

Theorem 1. A strategy (Π,Q,R) is ADU for a if and only if $(I-R\Pi)a = Q\Pi x$ for some $x \in M(X)$.

While a purely model-based viewpoint leads to the BLU linear projection strategies (Π,Σ^{-1},I), the imposition of asymptotic design unbiasedness favors generalized regression strategies (Π,Q,Π^{-1}). Since $0 \in M(X)$, Theorem 1 implies that a generalized regression strategy is ADU for all a. In fact, any strategy (Π,Q,R) is ADU for a if and only if it is equivalent to the generalized regression strategy (Π,Q,Π^{-1}) for a.

For this purpose, two strategies, (Π,Q_1,R_1) and (Π,Q_2,R_2) are said to be equivalent for a if and only if they produce identical estimates of a'y for all y and all samples with positive probability of occurrence. Given identical Q, two strategies are equivalent if and only if $a'(R_1-R_2)\Delta\hat{u} = 0$ for all s and all y. But, as in the proof of Theorem 1, this is true if and only if $(R_1-R_2)a = Qx$ for some $x \in M(X)$. However Theorem 1 shows that a strategy (Π,Q,R) is ADU for a if and only if $(\Pi^{-1}-R)a = Qx$, $x \in M(X)$. This proves

Theorem 2. A strategy (Π,Q,R) is ADU for a if and only if (Π,Q,R) and the generalized regression strategy (Π,Q,Π^{-1}) are equivalent for a.

Several special cases may illustrate the utility of these results.

(a) The ratio model $k = 1$ with $x_I > 0$ is of great practical importance and has been intensively studied. Theorem 1 shows that (Π,Q,R) is ADU for a if and only if

$$q_I = (\lambda\pi_I x_I)^{-1}(1-r_I\pi_I)a_I. \tag{3.4}$$

Here $\lambda > 0$ is an arbitrary constant of proportionality. For a linear prediction strategy (Π,Q,I), (3.4) gives Brewer's (1979) relationship,

$$q_I = (\lambda x_I)^{-1}(\pi_I^{-1}-1)a_I, \quad \lambda > 0.$$

A simple projection strategy $(\Pi,Q,0)$ is ADU if and only if $q_I = (\lambda\pi_I x_I)^{-1}a_I$, giving the estimator

$$a'\hat{y} = \Big(\sum_{J\in s} \pi_J^{-1} a_I y_I \Big/ \sum_{I\in s} \pi_I^{-1} a_I x_I\Big) \sum_{I=1}^{N} a_I x_I .$$

(b) (2.1) is said to include an intercept if $e = [1 \ldots 1]' \in M(X)$. In this case an ADU strategy for a can be constructed using Theorem 1 with $x = e$, giving $q_I = (\lambda\pi_I)^{-1}(1-r_I\pi_I)a_I$. An ADU linear prediction strategy is obtained using $q_I = \lambda^{-1}(\pi_I^{-1}-1)a_I$ while, for an ADU simple projection strategy, use $q_I = (\lambda\pi_I)^{-1}a_I$.

In the previous cases Q involves both Π and a, but sometimes ΠI or BLU strategies may be constructed that are ADU for a.

(c) (2.1) is said to be directed to a if $a \in M(X)$. In this case the strategies $(\Pi,\Pi^{-1},0)$ and $(\Pi,\Pi^{-1}-I,I)$ are ADU for a.

(d) A BLU strategy (Π,Σ^{-1},R) is ADU for a if and only if $\Pi^{-1}\Sigma(I-R\Pi)a \in M(X)$. In particular, a BLU linear prediction strategy (Π,Σ^{-1},I) is ADU for a if and only if $\Sigma(\Pi^{-1}-I)a \in M(X)$. This odd requirement seems to reflect the dissatisfaction of many samplers with these strategies. A somewhat nicer condition characterizes a BLU simple projection strategy, namely $\Pi^{-1}\Sigma a \in M(X)$.

4. EFFICIENCY OF ADU STRATEGIES

Within the class of ADU strategies, a useful planning criterion is the asymptotic variance of $a'\hat{y}$, denoted $v(a'\hat{y})$. Here $v(a'\hat{y})$ is defined to be the asymptotic expectation, with respect to both design and model, of the mean square prediction error of $a'\hat{y}$. The asymptotic construction is as developed in Section 3 but with y_j independently generated following (2.1). In this case, there are m independent u_j, with $E_\xi(u_j) = 0$ and $E_\xi(u_j u_j') = \Sigma$, $j = 1,\ldots,m$.

To examine the square error $(a'y_m - a'\hat{y}_m)^2$, use (3.2) to note that

$$\sum_{j=1}^{m} y_j - \hat{y}_j = \sum_{j=1}^{m} (I-\hat{T}_m\Delta_j)u_j, \text{ since}$$

$$\sum_{j=1}^{m} (I-\hat{T}_m\Delta_j)X = m(I-\hat{T}_m\hat{\Pi}_m)X$$

$$= m(I-R\hat{\Pi}_m)(I-\hat{C}_m\hat{\Pi}_m)X$$

$$= 0.$$

The ξ-independence of the u_j implies

$$E_\xi(a'y_m-a'\hat{y}_m)^2 = m^{-1}\, E_\xi[\sum_{j=1}^{m} a'(y_j-\hat{y}_j)]^2$$

$$= m^{-1} \sum_{j=1}^{m} a'(I-\hat{T}_m\Delta_j)\ \Sigma\ (I-\hat{T}_m'\Delta_j)a$$

$$= a'(I-\hat{\Pi}_m)\Sigma a + a'(I-\hat{T}_m)\hat{\Pi}_m\Sigma(I-\hat{T}_m')a.$$

Now the asymptotic design-based expectation can be evaluated as in Section 3, giving

$$\lim_{m\to\infty} E_d E_\xi(a'y_m-a'\hat{y}_m)^2 = a'(I-\Pi)\Sigma a + a'(I-T)\Pi\Sigma(I-T')a. \quad (4.1)$$

Given that (Π,Q,R) is ADU for a, $a'T\Pi y = a'y$ for all $y \in R^N$ so that (4.1) simplifies to $a'(\Pi^{-1}-I)\Sigma a$. This justifies

Definition 2. If (Π,Q,R) is ADU for a, then the asymptotic variance of (Π,Q,R) for a is

$$v(a'\hat{y}) = a'(\Pi^{-1}-I)\Sigma a \quad (4.2)$$

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

$v(a'\hat{y})$ becomes especially recognizable with an equal probability sample plan, $\pi_I = n/N$. In this case,

$$v(a'\hat{y}) = \frac{N^2}{n}\left(\frac{N-n}{N}\right)\left(\frac{1}{N}\sum_{I=1}^{N} a_I^2\,\sigma_I^2\right). \quad (4.3)$$

By defining the population variance of ay to be

$$\sigma_{ay}^2 = N^{-1} \sum_{I=1}^{N} a_I^2 y_I^2 - (N^{-1} \sum_{I=1}^{N} a_I y_I)^2,$$

and the coefficient of determination of (2.1) for ay to be

$$R_{ay}^2 = (\sigma_{ay}^2 - N^{-1} \sum_{I=1}^{N} a_I^2 \sigma_I^2)/\sigma_{ay}^2,$$

then

$$v(a'\hat{y}) = \frac{N^2}{n} (\frac{N-n}{N})(1-R_{ay}^2)\sigma_{ay}^2. \tag{4.4}$$

There are three ways of increasing the asymptotic precision of an ADU regression estimator:

(1) increase n,

(2) increase R_{ay}^2 by utilizing more relevant auxiliary information, and

(3) choose a more efficient strategy (Π,Q,R).

We now explore the latter possibility.

(4.2) shows that $v(a'\hat{y})$ depends only on Π for any ADU strategy (Π,Q,R). The Cauchy-Schwartz inequality implies that

$$(\sum_{I=1}^{N} |a_I|\sigma_I)^2 \leq (\sum_{I=1}^{N} \pi_I)(\sum_{I=1}^{N} a_I^2 \sigma_I^2 \pi_I^{-1})$$

with equality if and only if $\pi_I^{1/2}$ is proportional to $|a_I|\sigma_I\pi_I^{-1/2}$. Since $\sum_{I=1}^{N} \pi_I = n$, we have

Theorem 3. Within the class of strategies (π,Q,R) that are ADU for a and have sample size n, the minimum asymptotic variance is

$$v(a'y) = n^{-1}(\sum_{I=1}^{N} |a_I|\sigma_I)^2 - \sum_{I=1}^{N} a_I^2 \sigma_I^2.$$

The minimum asymptotic variance is achieved by an ADU strategy for a if and only if

$$\pi_I = n|a_I|\sigma_I \Big/ \sum_{J=1}^{N} |a_J|\sigma_J.$$

It is perhaps appropriate to call a strategy best for a if it is ADU for a and achieves the minimum asymptotic variance. If $|a_I|\sigma_I$ is called the relevance of I, then a strategy is best if and only if π_I is proportional to the relevance of I.

The best strategy depends very strongly on the population characteristic a. For any best strategy, $\pi_I = 0$ if and only if $a_I = 0$, so units not relevant to a are not sampled. A single strategy can only be best for two characteristics a and a* if |a| and |a*| are proportional. For example a strategy that is best for the population total, a = e, is also best for all differences between complementary subclass totals.

In the previous section, Theorem 1 was used to examine conditions allowing the construction of certain types of ADU strategies. The class of BLU simple projection strategies $(\Pi, \Sigma^{-1}, 0)$ seems especially appealing when the sample size is not large, and the model (2.1) is credible. Theorem 3 shows that there exists a best BLU simple projection strategy for a if and only if $\text{sign}(a_I)\ \sigma_I = x_I'\lambda$, for some $\lambda \in R^k$. In particular, a best BLU simple projection strategy exists for the population total if and only if $(\sigma_1 \ldots \sigma_N)' \in M(X)$.

Ordinarily it will not be possible to follow a strategy that is best for all a of interest. So it is useful to define the asymptotic efficiency for a of any strategy (Π,Q,R) that is ADU for a. Let n be the sample size of (Π,Q,R) and let $v(a'\hat{y})$ be its asymptotic variance. Suppose n* is the sample size of another strategy that is best for a and has the same asymptotic variance $v(a'\hat{y})$. Then it is natural to regard n*/n as the efficiency of (Π,Q,R) for a. But Theorem 3 implies that n*/n is equal to

$$(n \sum_{I=1}^{N} \pi_I^{-1} a_I^2 \sigma_I^2)^{-1} (\sum_{I=1}^{N} |a_I| \sigma_I)^2. \qquad (4.5)$$

This quantity is defined to be the asymptotic efficiency of (Π,Q,R) for a.

In certain cases, (4.5) is determined by the population coefficient of variation of the relevance $|a_I|\sigma_I$, denoted $cv_{a\sigma}$. For instance, consider an ADU strategy with inclusion probabilities proportional to $a_I^2 \sigma_I^2$. The efficiency of any such strategy is equal to

$$(N \sum_{I=1}^{N} a_I^2 \sigma_I^2)^{-1} (\sum_{I=1}^{N} |a_I| \sigma_I)^2$$

which is simply $(1 + cv_{a\sigma}^2)^{-1}$. An example of this is any ADU strategy for the population total with pps sampling, if σ_I^2 is proportional to size.

A second case, of greater interest, is any ADU strategy using an equal probability sampling plan, $\pi_I = n/N$. These strategies will usually be preferred in practice unless their efficiency is very poor. (4.5) shows that their efficiency is $(1 + cv_{a\sigma}^2)^{-1}$. This means that an equal probability sampling plan will be reasonably efficient if and only if all units are more or less equally relevant. Such a plan will be reasonably efficient for the population total if and only if (2.1) is reasonably homoscedastic. However, experience suggests that in many applications the relevant coefficient of variation is well in excess of unity, so that an equal probability sampling plan often has efficiency below 50% even for the population total. Such cases may call for a sampling plan providing inclusion probabilities more in line with the relevance of units.

5. STRONGLY STRATIFIED STRATEGIES

We now consider sample design in situations in which the efficiency of an equal probability ADU strategy is poor enough to justify the use of unequal π_I. In these cases, stratification can provide nearly optimal inclusion

probabilities giving strategies which are simple to execute and very compatible with common practice. By using equal π_I within each stratum, these stratified sampling strategies side-step most of the problems that are encountered with general varying probability designs. Moreover there is no significant loss in efficiency.

Suppose that $\{S_h: h=1,\ldots,H\}$ is any stratification of the population, and let cv_h be the coefficient of variation of $|a_I|\sigma_I$ within the N_h units of stratum h, so that

$$1 + cv_h^2 = N_h \sum_{I\epsilon S_h} a_I^2 \sigma_I^2 \Big(\sum_{I\epsilon S_h} |a_I|\sigma_I \Big)^{-2}. \qquad (5.1)$$

We are interested in stratifications satisfying

$$cv_h \leq \epsilon, \quad h = 1,\ldots,H \qquad (5.2)$$

for some specified small $\epsilon > 0$. For any such stratification, we consider a sampling plan having sample allocation proportional to the aggregate relevance of units within each stratum, i.e. with $\pi_I = n_h/n$, for $I \in S_h$, where

$$n_h = n\Big(\sum_{I\epsilon S_h} |a_I|\sigma_I \Big) \Big/ \sum_{I=1}^{N} |a_I|\sigma_I, \quad h = 1,\ldots,H. \qquad (5.3)$$

Equivalently the sampling fractions n_h/N_h are proportional to the average relevance of units within each stratum.

Definition 3. A strategy (Π,Q,R) is strongly stratified for a if

a) the stratification satisfies (5.2) for a specified small ϵ,

b) the allocation follows (5.3), and

c) Q and R define an estimator $a'\hat{y}$ which is ADU for $a'y$.

Theorem 3 provides a lower bound on the asymptotic variance of any ADU strategy with sample size n. However, for any strongly stratified strategy a tight upper bound is also easily derived from (4.2), (5.1) and (5.3):

$$v(a'\hat{y}) = \sum_{h=1}^{H} (N_h/n_h - 1) \sum_{I\epsilon S_h} a_I^2 \sigma_I^2$$

$$= n^{-1} \sum_{I=1}^{N} |a_I| \sigma_I \sum_{h=1}^{H} (1+cv_h^2) \sum_{I \epsilon S_h} |a_I| \sigma_I - \sum_{I=1}^{N} a_I^2 \sigma_I^2$$

$$\leq (1+\epsilon^2) n^{-1} \left(\sum_{I=1}^{N} |a_I| \sigma_I \right)^2 - \sum_{I=1}^{N} a_I^2 \sigma_I^2. \qquad (5.4)$$

Equivalently, we have

Theorem 4. The asymptotic efficiency of a strongly stratified strategy is at least $(1+\epsilon^2)^{-1}$.

The allocation rule (5.3) is not actually optimal in terms of minimizing $v(a'\hat{y})$. The optimal allocation is to choose n_h proportional to $(N_h \sum_{I \epsilon S_h} a_I^2 \sigma_I^2)^{1/2}$, giving

$$v(a'\hat{y}) = n^{-1} \left[\sum_{h=1}^{H} (1+cv_h^2)^{1/2} \sum_{I \epsilon S_h} |a_I| \sigma_I \right]^2 - \sum_{I=1}^{N} \sigma_I^2.$$

However, as long as ϵ is small this cannot be much better than the simpler allocation (5.3).

As long as ϵ is small, all strongly stratified strategies are almost equivalent in terms of asymptotic efficiency, so the actual choice of stratification is almost inconsequential. The complexities of optimal stratification (discussed in Anderson, Kish and Cornell, 1980; Rao, 1977; and Singh, 1971, 1975) can be avoided simply by utilizing a regression estimator, so that the efficiency of the design depends only on the residual variation and not on the within-strata variation of the auxiliary variables. Any convenient construction of strata can be used, as long as the ϵ is small, including the Dalenius-Hodges procedure. In practice it may be advantageous to use a design with equal n_h. Following (5.3) this is achieved by constructing strata to equalize the aggregate within-strata relevance of units, i.e. by equalizing $\sum_{I \epsilon S_h} |a_I| \sigma_I$. This is a generalization of the equal aggregate size recommendation of Hanson, Hurwitz and Madow (1953, p. 219). Cochran (1961) seems to discredit this simple rule, but his findings are due to his failure to use

the available auxiliary information not only in the sampling plan but also in the estimator, i.e. in both components of the sampling strategy.

6. APPLICATIONS

The methods of sample design proposed in this paper are relevant whenever a key target variable is to be measured on a sampling basis from a frame which provides one or more relevant predictor variables. Although this situation is encountered in a variety of contexts, the two examples to be discussed involve accounting for energy usage.

In electric utility load research, the target variable y is often customer consumption (i.e. "demand") of electricity during certain peak hours, and the characteristic of interest is the population total of y. The auxiliary information in the simplest case is the monthly usage of electricity (x) that is metered for billing each customer. Analysis can usually be based on a simple heteroskedastic ratio model relating peak period demand to monthly usage. The analysis illustrated by this example is applicable in most sampling situations in which the univariate ratio estimator would ordinarily be used.

In the second example the population is comprised of 205 buildings operated by a major university, and the target variable y is the heating cost of each building, which is related to several measures of the size and usage of the building. The characteristic of interest $a'y$ is the share of total heating costs that can be allocated to sponsored research. The vector a is considered to be known from space usage reports for each building, but y is only available on a sampling basis. This example will illustrate the use of multivariate auxiliary information and a nontrivial characteristic of interest. This example also illustrates sample design with 100% inclusion of the most relevant units.

6.1 A Load Research Example

This example is based on a dataset that Brandenburg and Higgins (1974) have previously used to illustrate sample design for load research. The dataset provides peak demand y_I (in kw) and monthly usage (in mwh) for each of $n = 210$ commercial and industrial customers. We will use these data, called the analysis sample, to plan a new sample of a population of $N = 840$ customers with known x but unknown y. The purpose of the new sample is to estimate (or predict) $\sum_{I=1}^{N} y_I$.

To plan the new sample, the analysis sample will be used to estimate the parameters of a superpopulation model (2.1) that is assumed to underlie both the analysis sample and the target population. It sometimes may be useful to pool the data from several available past studies and possibly to take into account trends or other changes in superpopulation parameters, but these complexities will not be introduced here. However planning will take full account of the known distribution of x in the target population.

Figure 6.1 shows a scatterplot of the analysis sample. Exploratory analysis and experience with several other load research datasets suggest the simple heteroskedastic ratio model:

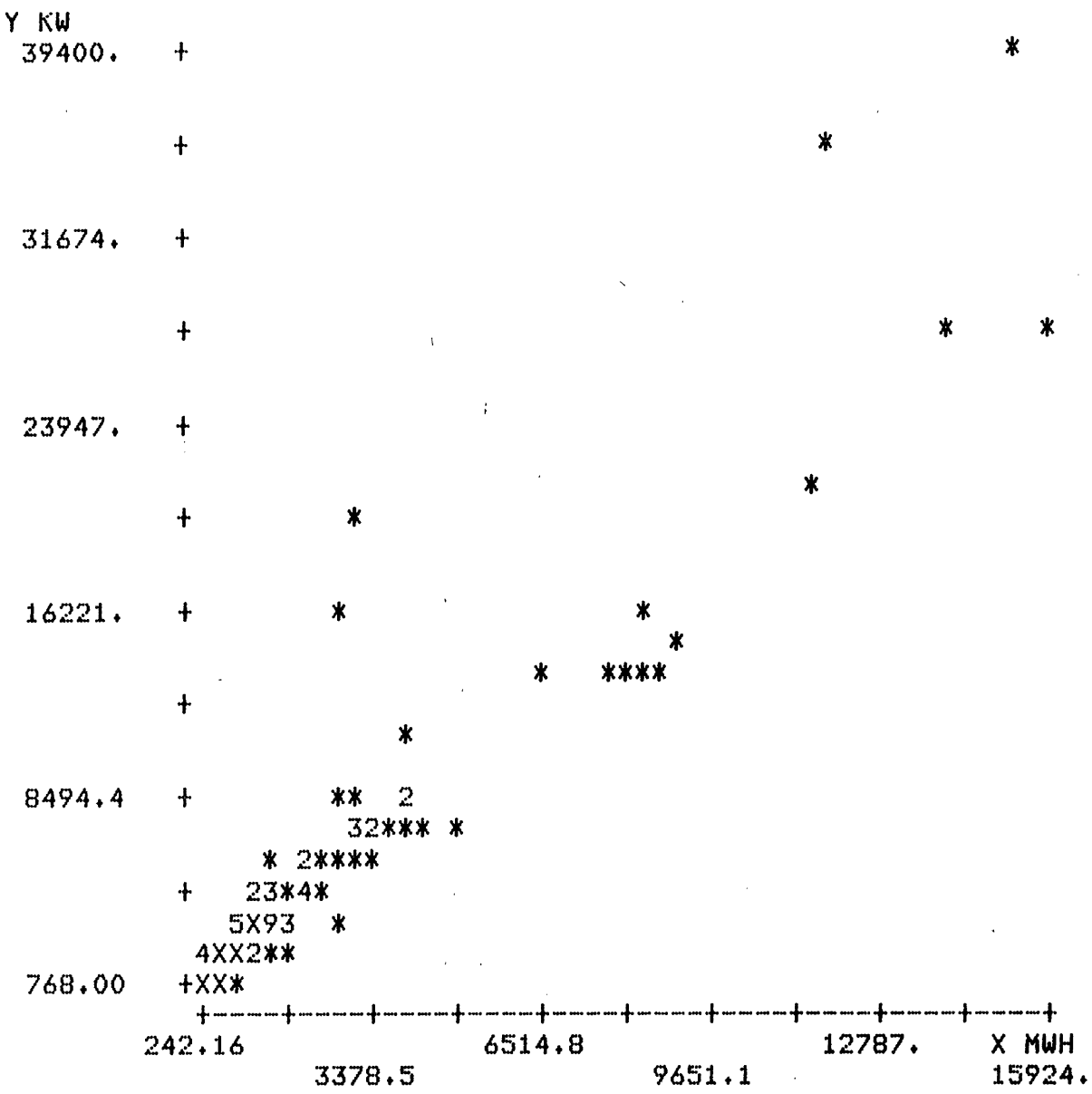
$$y_I = \beta x_I + u_I, \text{ with} \qquad (6.1)$$

$$\sigma_I = \sigma_0 \, x_I^{\gamma}.$$

It is further assumed that u_I is normally distributed, although there are two of observations that seem to strain this assumption.

The assumed normality can be used to calculate model-based maximum likelihood estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\sigma}_0$ using an iterative algorithm (Harvey , 1976).

Figure 6.1 Scatterplot of Load Research Data



To describe the algorithm, consider the more general model $y_I = X_I'\beta + u_I$ with $\sigma_I = \sigma_0 z_I^{\gamma}$, $z_I > 0$. Conditional on an initial estimate $\hat{\gamma}_0$, weighted least squares gives

$$\hat{\beta} = \Big(\sum_{I\epsilon s} z_I^{-2\hat{\gamma}_0} X_I X_I'\Big)^{-1} \sum_{I\epsilon s} z_I^{-2\hat{\gamma}_0} X_I y_I, \qquad (6.2)$$

$$\hat{\sigma}_0^2 = n^{-1} \sum_{I\epsilon s} \hat{v}_I^2, \text{ where}$$

$$\hat{v}_I = z_I^{-\hat{\gamma}_0} (y_I - X_I'\hat{\beta}).$$

A revised estimate $\hat{\gamma}_1 = \hat{\gamma}_0 + \Delta\hat{\gamma}$ is obtained by calculating the ordinary least squares regression coefficient

$$\Delta\hat{\gamma} = \Big(\sum_{I\epsilon s} w_I^2\Big)^{-1} \sum_{I\epsilon s} w_I c_I, \text{ where} \qquad (6.3)$$

$$w_I = \hat{v}_I^2 / 2\hat{\sigma}_0^2, \text{ and}$$

$$c_I = \log(z_I) - n^{-1} \sum_{I\epsilon s} \log(z_I).$$

This is repeated until convergence.

With the analysis dataset, this algorithm gives the estimated relationships

$$\hat{y}_I = 2.737 x_I \text{ and}$$

$$\hat{\sigma}_I = .9223 x_I^{.9832}.$$

Although the distribution of x in the target population was not published, the following target population statistics are consistent with the analysis sample:

$$N = 840$$

$$N^{-1} \sum_{I=1}^{N} \hat{y}_I = 4353.9 \text{ kw},$$

$$N^{-1} \sum_{I=1}^{N} \hat{\sigma}_I = 1278.9 \text{ kw, and}$$

$$N^{-1} \Big(\sum_{I=1}^{N} \hat{\sigma}_I^2\Big)^{1/2} = 2322.6 \text{ kw}.$$

These statistics, together with the partial sums of $\hat{\sigma}$ and $\hat{\sigma}^2$ with cases in order of increasing $\hat{\sigma}$, are all that are needed to develop an efficient sampling plan.

The first step of analysis is to calculate the sample size that would be required to estimate $a'y = \sum_{I=1}^{N} y_I$ using an ADU estimator with an equal probability sampling plan. PURPA specifies $\pm$ 10% or less relative error with 90% probability. Using this criterion, the asymptotic variance v(a'y), given by (4.3), should satisfy

$$1.645 \sqrt{v(a'y)}/a'y = .10,$$

or equivalently,

$$1.645 \frac{1}{\sqrt{n}} \sqrt{\frac{N-n}{n}} \, cv_{au} = .10.$$

The statistic cv_{au}, called the residual coefficient of variation of relevance, is

$$cv_{au} = (N^{-1} \sum_{I=1}^{N} a_I^2 \sigma_I^2)^{1/2}/(N^{-1} \sum_{I=1}^{N} a_I y_I) \qquad (6.4)$$

$$\doteq (N^{-1} \sum_{I=1}^{N} \hat{\sigma}_I^2)^{1/2}/(N^{-1} \sum_{I=1}^{N} \hat{y}_I)$$

$$= .5335.$$

When the sampling fraction n/N is negligible, the sample size required with equal probability sampling is

$$n_0 = (1.645 \, cv_{au}/.10)^2$$

$$= 77 \text{ customers.}$$

Correcting for the finite population, the sample size required with equal probability sampling is

$$n_1 = n_0/(1+n_0/N)$$

$$= 70.53 \text{ customers.}$$

The second step of analysis is to examine the reduction in n_1 resulting from the best varying probability sampling plan. Using results of Section 4,

the asymptotic efficiency of an equal probability sampling plan relative to the best varying probability sampling plan is

$$\begin{aligned} \text{eff} &= (1+cv_{a\sigma}{}^2)^{-1} \\ &= (N^{-1}\sum_{I=1}^{N}|a_I|\sigma_I)^2/(N^{-1}\sum_{I=1}^{N}a_I{}^2\sigma_I{}^2) \\ &\doteq (N^{-1}\sum_{I=1}^{N}\hat{\sigma}_I)^2/(N^{-1}\sum_{I=1}^{N}\hat{\sigma}_I{}^2) \\ &\doteq .3032. \end{aligned}$$

This means that the best varying probability sampling plan will require a sample size of

$$\begin{aligned} n_2 &= \text{eff}\cdot n_1 \\ &\doteq 21.38 \text{ customers.} \end{aligned}$$

In practice, n might be increased to thirty customers to increase confidence in the accuracy of the asymptotic approximations.

The best varying probability sampling plan uses inclusion probabilities proportional to $\hat{\sigma}_I$ or $x_I{}^{\hat{\gamma}}$. Since $\hat{\gamma}$ is so close to one, an alternative design would be to use inclusion probabilities proportional to x_I, i.e. to use a pps sampling plan. Using (4.5) and the additional target population statistic

$$N^{-1}\sum_{I=1}^{N}\hat{\sigma}_I{}^2/x_I = 1029.7,$$

the asymptotic efficiency of the pps plan is about

$$\begin{aligned} &(N^{-1}\sum_{I=1}^{N}\hat{\sigma}_I)^2/(N^{-1}\sum_{I=1}^{N}x_I \cdot N^{-1}\sum_{I=1}^{N}\hat{\sigma}_I{}^2/x_I) \\ &= .9996. \end{aligned}$$

Another alternative is to use a strongly stratified sampling plan along the lines of Section 5. For any strong stratification, the sample is allocated among strata in proportion to the within-strata totals of $\hat{\sigma}_I$. These

strata can be defined in any convenient fashion provided only that the coefficient of variation of $\hat{\sigma}_I$ is small within all strata. In particular, a balanced design can be obtained by dividing the aggregate target population total of $\hat{\sigma}_I$ about equally among strata. This is easily done by examining the cumulative sum of the $\hat{\sigma}_I$ in increasing order.

Table 6.1 showed a stratified sampling design using six strata with five observations per strata, designed to equalize $\sum_{I \epsilon S_h} \hat{\sigma}_I$ as much as possible. The last column shows that the efficiency of this design is at least .92. In fact, this is somewhat conservative, since (4.5) gives the asymptotic efficiency as .95.

Table 6.1 Strongly Stratified Design for Load Research Example

Stratum h	Size N_h	Upper Boundary x_h mwh	Sampling Fraction n_h/N_h %	$\sum_{I \epsilon S_h} \hat{\sigma}_I$ 10^4kw	$\sum_{I \epsilon S_h} \hat{\sigma}_I^2$ 10^8kw^2	$(1+cv_h^2)^{-1}$
1	440	728	1.1	17.83	.7764	.93
2	204	1,554	2.5	17.76	1.638	.94
3	100	3,042	5.0	17.82	3.329	.95
4	56	7,955	8.9	19.05	7.032	.92
5	24	11,596	20.8	17.12	12.40	.99
6	16	16,000	31.3	17.84	20.14	.99

This simple stratified sampling plan sidesteps the controversy involved in estimating the achieved precision of the pps plan. With the ratio model and stratified sampling, the simple projection strategy discussed in Section 3 gives the combined ratio estimator. In this case, the expressions for the expected asymptotic variance are closely related to the traditional design-based

measure of precision for the combined ratio estimator. So, while planning is necessarily model-based, post-sampling analysis can be conventional if desired.

Table 6.2 summarizes this sort of analysis for twelve different load-study populations. In the first five of these examples, an equal probability sampling plan can be teamed with the ordinary ratio estimator to provide a rather efficient strategy. In fact each of these is a population of residential customers. The remaining populations, in which varying probability designs will be more advantageous, are all groups of commercial, industrial or municipal customers which are characterized by high variation in x and strong heteroskedasticity in the relationship between y and x.

The wide variation of $\hat{\gamma}$, eff, and n_2 in these examples dramatizes the need to tailor a sampling strategy to the characteristics of each population. Of course some variation in these statistics is due to their sampling distributions, but simulation experiments indicate that these statistics are rather reliable. Simulation can also be used to explore the validity of the asymptotic approximations to the mean and variance of $a'\hat{y}$. These results, to be reported in a later paper, are favorable although very small samples, say less than thirty, are not generally recommended.

6.2 A Cost Accounting Example

One purpose of utility load research is to allocate the indirect cost of maintaining system capacity in proportion to the peak demands of various subclasses of customers. The second example involves a related problem of cost allocation. In this example, the administration of a large university wants to allocate some of its heating costs to sponsored research. For each of its $N = 205$ buildings, the administration knows the proportion $a_I \geq 0$ of the building assignable to sponsored research. However, the cost of heating

each building (y_I) is unknown, although the total cost $\sum_{I=1}^{N} y_I$ is known. The characteristic of interest is $a'y = \sum_{I=1}^{N} a_I y_I$.

Table 6.2 Statistics of Other Rate Research Populations

Example	Equal Probability n_1	$\hat{\gamma}$	Standard Error of $\hat{\gamma}$	Efficiency	Best n_2	Analysis Sample Size
1	134	.51	.29	.92	123	185
2	12	.62	.23	.88	10	29
3	38	1.08	.41	.81	31	30
4	17	.86	.18	.77	13	51
5	166	.91	.14	.75	125	185
6	44	.89	.14	.57	25	32
7	40	.84	.12	.57	23	32
8	92	.77	.06	.44	41	73
9	32	.74	.07	.38	12	83
10	684	1.05	.09	.29	198	90
11	983	1.38	.11	.21	205	63
12	5,621	1.28	.08	.16	876	30

A simple ratio model might be used to relate y_I to the total size of each building x_I (measured in square feet):

$$y_I = \beta x_I + u_I.$$

In this model, the expected heating cost per square foot is identical for all buildings and is the coefficient β. With this assumption, sampling is unnecessary since β can be estimated as

$$\hat{\beta} = \sum_{I=1}^{N} y_I / \sum_{I=1}^{N} x_I$$

and a'y can be estimated as a'$\hat{y}$ where $\hat{y}_I = \hat{\beta}x_I$. However this assumption is unrealistic. For example, heavy use of fume hoods in chemical laboratories significantly increases heating costs since replacement air must be heated.

A more realistic model relates heating costs to several categories of building use. Define a vector $X_I = (x_{I1}, \ldots, x_{Ik})' \in R^k$ with x_{Ij} equal to the square footage of building I in use category j. Then the model $y_I = X_I'\beta + u_I$ introduces a vector $\beta \in R^k$ of coefficients associated with the distinct use categories. In this case y_I can be measured for a sample of buildings and this sample data can be used to estimate β and a'y.

If the cost of measuring y_I is high, it is worthwhile to develop an efficient sampling strategy. In order to develop a strategy, an analysis database has been assembled which includes the known a_I and X_I and a preliminary estimate of y_I for all 205 buildings. Three use categories have been used:

1. General, including classrooms and offices;
2. Laboratory, both class and research; and
3. Nonassignable, including out-of-use, custodial, and structural areas.

Analysis of these data using the algorithm (6.2)-(6.3) led to the estimated relationship

$$\hat{y}_I = 0.371\, x_{I1} + 2.359\, x_{I2} + 2.359\, x_{I3}, \text{ with}$$

$$\hat{\sigma}_I = 22.43\, (x_{I2} + x_{I3})^{.7594}$$

and to the following finite population statistics:

$$N = 205$$

$$N^{-1} \sum_{I=1}^{N} a_I \hat{y}_I = \$17,520$$

$$N^{-1} \sum_{I=1}^{N} a_I \hat{\sigma}_I = \$10,614$$

$$(N^{-1} \sum_{I=1}^{N} a_I^2 \hat{\sigma}_I^2)^{1/2} = \$26,439.$$

Initially, the analysis follows the same steps as the previous example. If an error limit of $\pm$ 10% with 95% probability is adopted, the required sample size using an equal probability sampling plan uncorrected for the finite population is

$$n_0 = (1.96\ cv_{au}/.10)^2$$

$$= 874.8 \text{ buildings,}$$

where

$$cv_{au} \doteq (N^{-1} \sum_{I=1}^{N} a_I^2 \hat{\sigma}_I^2)^{1/2} / (N^{-1} \sum_{I=1}^{N} a_I \hat{y}_I)$$

$$= 1.509.$$

This is corrected for the finite population size N = 205:

$$n_1 = n_0/(1+n_0/N)$$

$$= 166.1 \text{ buildings.}$$

The efficiency of this equal probability sampling plan is

$$eff = (1+cv_{a\sigma}^2)^{-1}$$

$$\doteq (\sum_{I=1}^{N} a_I \hat{\sigma}_I)^2 / \sum_{I=1}^{N} a_I^2 \hat{\sigma}_I^2$$

$$= .1612,$$

so that this plan can be greatly improved.

An alternative plan is to select buildings with probability proportional to their size as measured by $x_2 + x_3$. Using (4.5), the efficiency of this

plan is approximately

$$\Big(\sum_{I=1}^{N} a_I\hat{\sigma}_I\Big)^2 \Big/ \Big[\sum_{I=1}^{N}(x_{I2}+x_{I3}) \sum_{I=1}^{N} a_I^2\hat{\sigma}_I^2(x_{I2}+x_{I3})^{-1}\Big]$$

$$= .2616;$$

so this helps, but not much.

The best plan, in the sense of Theorem 3, is to select units with probability proportional to their relevance for the characteristic of interest, i.e. with probability proportional to $a_I\hat{\sigma}_I$. The sample size n_2 that would be required with this plan can be calculated from the size and efficiency of the equal probability plan:

$$n_2 = (\text{eff})n_1$$

$$= 26.78 \text{ buildings.}$$

This figure n_2 should be regarded as a lower bound that can only be achieved by using the optimal inclusion probabilities of Theorem 3. However in this case, $n_2|a_I|\hat{\sigma}_I / \sum_{J=1}^{N}|a_J|\hat{\sigma}_J$ exceeds one for the most relevant units, so the optimal rule is infeasible. In this situation the best feasible design is to use 100% sampling for units M+1 to N with optimal choice of π_I for $I \leq M$. Here the units are considered to be in order of increasing relevance, and M is found as follows. Let v be the required value of v(a'y):

$$v = n^{-1}\Big(\sum_{I=1}^{M}|a_I|\hat{\sigma}_I\Big)^2 - \sum_{I=1}^{M} a_I^2\hat{\sigma}_I^2$$

since units M+1, ..., N contribute no variance. Moreover, for I=1, ..., M,

$$\pi_I = n|a_I|\hat{\sigma}_I / \sum_{J=1}^{M}|a_J|\hat{\sigma}_J$$

$$= |a_I|\hat{\sigma}_I \sum_{J=1}^{M}|a_J|\hat{\sigma}_J/\Big(\sum_{J=1}^{M} a_J^2\hat{\sigma}_J^2 + v\Big).$$

In particular, M is the largest unit such that

$$\pi_M = |a_M|\hat{\sigma}_M \sum_{J=1}^{M} |a_J|\hat{\sigma}_J / \left(\sum_{J=1}^{M} a_J^2 \hat{\sigma}_J^2 + v \right) \qquad (6.5)$$

$$< 1.$$

Using $v = (.10 \sum a_I \hat{y}_I / 1.96)^2$

$$= 3.35787 \times 10^{10},$$

M turns out to be 193. So the best feasible design is to select the 12 most relevant units with certainty, and to select n_3 additional units with probability proportional to relevance as in Theorem 3. Here

$$n_3 = \left(\sum_{I=1}^{M} |a_I|\hat{\sigma}_I \right)^2 / \left(\sum_{I=1}^{M} a_I^2 \hat{\sigma}_I^2 + v \right)$$

$$= 16.43 \text{ buildings.}$$

This may be raised to 18 to comply with the convention of using a sample of at least 30 observations. In fact, 123 buildings are totally irrelevant in the sense that $a_I = 0$, so these 18 buildings are selected from a very small population of 70 buildings.

It may be convenient to select these eighteen buildings using a stratified sampling plan. Table 6.3 summarizes a preliminary plan with three buildings selected from each of six strata with approximately equal aggregate relevance in each stratum. The asymptotic efficiency of this plan is .87.

The last column of Table 6.3 shows that the inefficiency comes mostly from stratum one. Table 6.4 shows a subdivision of stratum one into three strata with one sample building per stratum. With this refinement the asymptotic efficiency is improved to .95 so this stratified plan is almost optimal.

All of this analysis relies on asymptotic approximations which require validation in this and any other application involving small or moderate sample sizes. In specific cases, both bias and mean squared error can be effectively examined through computer simulation of both the finite population and the sample, conditional on an assumed superpopulation model. This

Table 6.3 A Strongly Stratified Design for the Cost Accounting Example

Stratum h	Size N_h	Sampling Fraction n_h/N_h, %	$\sum_{I \epsilon S_h} a_I \hat{\sigma}_I$ x10^4	$\sum_{I \epsilon S_h} a_I^2 \hat{\sigma}_I^2$ x10^8	$(1+cv_h^2)^{-1}$
1	43	7	17.285	13.093	.53
2	9	33	17.538	35.457	.96
3	7	43	19.775	56.484	.99
4	4	75	15.341	59.567	.99
5	4	75	19.943	99.999	.99
6	3	100	18.052	108.727	1.00

Table 6.4 A Substratification of Stratum 1

Stratum h	Size N_h	Sampling Fraction n_h/N_h, %	$\sum_{I \epsilon S_h} a_I \hat{\sigma}_I$ x10^4	$\sum_{I \epsilon S_h} a_I^2 \hat{\sigma}_I^2$ x10^8	$(1+cv_h^2)^{-1}$
1a	30	3	6.085	1.778	.69
1b	9	11	5.895	4.264	.91
1c	4	25	5.305	7.050	1.00

technique can be used to study both aspects of the strategy, sample design, and estimator. Within the accuracy of the asymptotic approximations, a large class of estimators is unbiased and equally efficient, but simulation may reveal important differences in the performance of these estimators with small and moderate samples. This work is underway.

7. SUMMARY AND CONCLUSIONS

Most work in sampling methodology has been directed to survey research, public health, and other fields where auxiliary information is limited, where the study is multipurpose, and where most of the collected information is qualitative. The present work is directed to management applications of sampling in that the study is narrowly focused on one or just a few quantitative variables that are closely related to detailed auxiliary information readily available in an administrative database. This relationship can be exploited to plan efficient data collection and analysis--in particular to determine the required sample size and to determine the most relevant units to be included in the sample on a random basis with varying inclusion probabilities. The optimal sample design can often be well approximated by a one-way stratified sampling plan.

These sampling plans are based on an assumed superpopulation model for the relationship between the target variable and the auxiliary information. However the proposed strategies utilize estimators which are more conventionally based on the sample design in the sense that they are asymptotically design unbiased even if the assumed model is misspecified. This provides a kind of robustness that is important in many sampling applications.

The proposed methodology for sample design is based entirely on asymptotic approximations which need to be investigated in specific applications involving

small or moderate samples. Simulation can perform this task and perhaps reveal differences in the small-sample distributions of estimators that are asymptotically equivalent in terms of their mean and variance. For larger samples, the generalized regression estimators are expected to perform well.

In many management applications, multivariate regression models are widely and effectively used for data analysis. This paper has offered an approach to data collection which ties directly into these models.

REFERENCES

Aigner, D. J., (1979), "Bayesian Analysis of Optimal Sample Size and a Best Decision Rule for Experiments in Direct Load Control," Journal of Econometrics, 9, 209-222.

_____, and J. A. Hausman, (1980), "Correcting for Truncation Bias in the Analysis of Experiments in Time-of-Day Pricing of Electricity." Bell Journal of Economics, 11, 131-142.

Anderson, D. W., L. Kish and R. G. Cornell, (1980), "On Stratification, Grouping, and Matching," Scandinavian Journal of Statistics, 7, 61-66.

Belsley, D. A., E. Kuh and R. E. Welsch, (1980), Regression Diagnostics, John Wiley & Sons, New York.

Brandenburg, L. and C. E. Higgins, Jr., (1974), "Stratified Random Sampling Methods for Class Load Surveys for Electric Utilities," Applied Statistics for Load Research, Vol. III, Association of Edison Illuminating Companies, New York.

Brewer, K. R. W., (1963), "Ratio Estimation in Finite Populations: Some Results Deductible From the Assumption of an Underlying Stochastic Process," Australian Journal of Statistics, 5, 93-105.

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

Cassel, C. M., C. E. Sarndal and J. H. Wretman, (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," Biometrika. 63, 615-620.

_____, (1977), "Foundation of Inference in Survey Sampling," John Wiley & Sons, New York.

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.

Ericson, W. A., (1969), "Subjective Bayesian Models in Sampling Finite Population, I," Journal of the Royal Statistical Association, B, 31, 195-234.

Fuller, W., (1975), "Regression Analysis for Sample Survey," Sankyha, 37, C. Pt. 3, 117-132.

Hansen, M. H., W. Hurwitz and W. G. Madow, (1953), Sample Survey Methods and Theory, Vol. 1, John Wiley & Sons, New York.

_____, W. G. Madow and B. J. Tepping, (1978), "On Inference and Estimation from Sample Surveys," Proceedings of the Survey Research Section, American Statistical Association, 82-107.

Harvey, A. C., (1976), "Estimating Regression Models with Multiplicative Heteroscedasticity," Econometrica, 44, 461-464.

Hocking, R. R., (1976), "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32, 1-49.

Holt, D. and T. M. F. Smith, (1979), "Post Stratifications," Journal of the Royal Statistical Society, A, 142, Part 1, 33-46.

Jonrup, H. and B. Rennermalm, (1976), "Regression Analysis in Samples from Finite Populations," Scandinavian Journal of Statistics, 3, 33-36.

Kish, L. and M. R. Frankel, (1974), "Inference from Complex Surveys," Journal of the Royal Statistical Society, B, 36, 1-37.

Konijn, H. S., (1962), "Regression Analysis in Sample Surveys," Journal of the American Statistical Association, 57, 590-606.

_____, (1973), Statistical Theory of Sample Survey Design and Analysis, North Holland, Amsterdam and American Elsevier, New York.

Raj, D., (1965), "On a Method of Using Multi-Auxiliary Information in Sample Surveys," Journal of the American Statistical Association, 60, 270-277.

Rao, C. R., (1973), Linear Statistical Inference and Its Applications, Second Edition, John Wiley & Sons, New York.

Rao, T. J., (1977), "Optimum Allocation of Sample Size and Prior Distributions: a Review," International Statistical Review, 45, 173-179.

Royall, R. M., (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, 377-387.

_____, (1971), "Linear Regression Models in Finite Population Sampling Theory," Foundations of Statistical Inference, V. P. Godambe and D. A. Spratt (eds.), Holt, Rinehart & Winston, Toronto.

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

_____, and J. Herson, (1973a), "Robust Estimation in Finite Populations," Journal of the American Statistical Association, 68, 880-889.

_____, (1973b), "Robust Estimation in Finite Population, II: Stratification on a Size Variables;" Journal of the American Statistical Association, 68, 891-893.

Sarndal, C. E., (1980), "On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling," to appear in Biometrika.

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.

_____, and T. M. F. Smith, (1969), "Estimation in Multistage Surveys," Journal of the American Statistical Association, 64, 830-840.

Singh, R., (1971), "Approximately Optimal Stratification on the Auxiliary Variable," Journal of the American Statistical Association, 66, 829-830.

_____, (1975), "On Optimal Stratification for Proportional Allocation," Sankhya, 37, C, Pt. 1, 109-115.

Smith, T. M. F., (1976), "The Foundations of Survey Sampling, A Review," Journal of the Royal Statistical Society, A, 139, Part 2, 183-204.

Taylor, L. D., (1977), "On Modeling the Residential Demand for Electricity by Time-of-Day," in Forecasting and Modeling Time-of-Day and Seasonal Electricity Demands, Electric Power Research Institute, Palo Alto, CA.