THE DESIGN CONSISTENT REGRESSION ESTIMATOR AND ITS CONDITIONAL VARIANCE

Phillip S. Kott, National Agricultural Statistics Service
Washington, DC 20250-2000

1. INTRODUCTION

There is a growing recognition, even among proponents of random sampling, that drawing inference about a finite population paramater without conditioning on known attributes of the sample is misleading. Articles advocating conditioning in a randomization framework include Rao (1985), Holt and Smith (1979), Oh and Scheuren (1983), and Hidiroglou and Sarndal (1986). In these, inferences are made - that is, expectations and variances are computed - with respect to a subset of samples with properties similar to the drawn sample. By contrast, the standard practice in design-based sampling theory is to take expectations over the set of all possible samples.

Royall (1983) would argue that even this type of conditioning may be "inferentially wrong" (see also 1970 and 1976). Rather than conditioning on a subset of possible samples, Royall assumes a model and then condition on the most relevant subset of all - the sample itself.

The problem with this purely model-dependent approach is that the model one assumes is almost always wrong, if only slightly. With this in mind, Fuller and Isaki (1981) reasonably suggested that, where possible, attention be limited to design consistent estimators. The design consistent regression estimators they proposed usually have small design biases but are model (conditionally) unbiased. Brewer (1979), Sarndal (1980), Robinson and Sarndal (1983), and Wright (1983) have made proposals along similar lines.

Advocates of combining design and model-based sampling theory usually focus their attention on design rather than model mean squared error. For example, see Fuller (1981), Sarndal (1982), and Wu (1982 and 1985). Design mean squared error is an important concept when designing a sampling plan. It has less inferential value once a sample is drawn however; at that point, statisticians and the users of their statistics should be more concerned with the accuracy of a realized estimate than with the "average" accuracy computed over all samples.

Wu and Deng (1983) put it this way: "the purpose of variance estimation is rather for assessing the variability of the ... estimator than for estimating the [design] variance itself."

Theoretical, model-dependent articles on conditional variance estimation include Royall and Eberhardt's investigation of the ratio estimator (1975), Royall and Cumberland on the general regression estimator (1978), Cumberland and Royall (1981), Royall (1986), and Valliant (1987). All are deeply concerned with a certain type of model failure - misspecification of the variance structure. Conditional variance estimators are proposed that are robust to this limited type of model failure in large samples with small sampling fractions. However, these articles do not satisfactorily address - as we will - the impact of more serious model failure resulting from missing and/or misspecified regressors.

Design consistent regression estimation for a population mean is reviewed in Section 2. Section 3 shows how a standard Yates-Grundy variance estimator can often be adjusted to be simultaneously a model unbiased estimator of the conditional (model) variance and a design consistent estimator of design mean squared error. Section 4 applies this general approach to some common sampling designs focusing particularly on probability weighted ratio estimators for which one popular conditional variance estimator is remarkably robust. Section 5 discusses some possible extentions.

2. DESIGN CONSISTENT REGRESSION ESTIMATION

2.1. Design Consistency

Suppose we have a population of size N. Each unit i in the population has associated with it a characteristic of interest, y_i. By drawing a sample of distinct units of size n<N, we would like to estimate the population mean $\bar{y}_N=\sum y_i/N$. Although we are unaware of the y_i values for units not chosen for the sample, we do know the values of a k element row vector of covariates, $\mathbf{x}_i$, for all units in the population.

Let p_i be the probability of choosing unit i for the sample, and let the units be re-arranged so that the sample consists of the units labeled 1, 2,..., n. Now consider estimators of the form

$$\hat{y}=[\mathbf{t}'\mathbf{y}_n+(\mathbf{1}_N'\mathbf{X}_N-\mathbf{t}'\mathbf{X}_n)\mathbf{b}]/N, \quad (1)$$

where $\mathbf{t} = (1/p_1,\ 1/p_2,\ \ldots,\ 1/p_n)'$;
$\mathbf{y}_m = (y_1,\ y_2,\ \ldots,\ y_m)'$, m=n or N;
$\mathbf{1}_m$ is an m-vector of 1's;
$\mathbf{X}_m$ is a m x k matrix whose ith row is $\mathbf{x}_i$;
and $\mathbf{b}$ is an as yet unspecified k-vector, which may be a function of the sampled y_i values.

Following Isaki and Fuller (1982), $\hat{y}$ is said to be <u>design consistent</u> when $\hat{y}-\bar{y}_N$ converges to zero in probability as the sample size, n, grows arbitrarily large (formally, a sequence of nested populations can be hypothesized so that n can become arbitrarily large).

Isaki and Fuller show that the following assumptions will force the design mean squared of $\hat{y}$ to be $O(n^{-1})$, which in turn will render $\hat{y}$ design consistent:

k, $|y_i|$, $|x_{ij}|$, and $|b_i|$, are all bounded for $i=1, \ldots, n$ and $j=1, \ldots, k$, (2a)

$$Np_i/n>M_1>0, \qquad (2b)$$

and $$\sum_i^N \sum_{j \neq i}^N h_{ij}/n<M_2 \qquad (2c)$$

where

$$h_{ij}=p_ip_j-p_{ij} \text{ when } p_ip_j-p_{ij} \text{ is positive}$$
$$=0 \quad \text{otherwise,}$$

and p_{ij} is the joint selection probability of units i and j.

The restrictions on the sampling design in (2b) and (2c) preclude at least one popular sampling plan: systematic sampling from a list with predetermined order. Kott (1986) showed why such a plan can not be part of a design consistent estimation strategy. Systematic sampling from a <u>randomly</u> ordered list, on the other hand, does satisfy the restrictions in (2b) and (2c).

2.2. <u>Model Unbiasedness</u>

The estimator in (1) was introduced with the folliwng linear regression model in mind:

$$\mathbf{y}_N=\mathbf{X}_N\boldsymbol{\beta}+\boldsymbol{\varepsilon}_N, \qquad (3)$$

where $E(\boldsymbol{\varepsilon}_N)=0$. It is easy to see that when $\mathbf{b}=\hat{\boldsymbol{\beta}}=\mathbf{Cy}_n$ for a for a k x n matrix $\mathbf{C}$ such that $\mathbf{CX}_n=\mathbf{I}_k$, the expectation of $\hat{y}-\bar{y}_N$ with respect to the random vector $\boldsymbol{\varepsilon}_N$ is zero.

We will label expectations with respect to $\boldsymbol{\varepsilon}_N$, E_ε, while expectations with respect to the sampling design will be denoted E_p. Variances will follow the same notation. When $E_\varepsilon(\hat{y}-\bar{y}_N)=0$, $\hat{y}$ is said to be <u>model unbiased</u>. Note that $\hat{y}$ remains design consistent even when the model in (3) fails as long as $\mathbf{b}=\mathbf{Cy}_n$ is bounded and the rest of equation (2) holds.

An example of a matrix $\mathbf{C}$ satisfying $\mathbf{CX}_n=\mathbf{I}_k$ is

$$\mathbf{C}_W=(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1},$$

where $\mathbf{W}$ is an n x n postive definite matrix (throughout the text, we assume that $\mathbf{X}_n$ is of full rank for convenience). If $\mathbf{W}=\mathbf{I}_n$ then $\hat{\boldsymbol{\beta}}$ is simply the ordinary least squares regression estimator of $\boldsymbol{\beta}$ in (3). In general, a design consistent, model unbiased $\hat{y}$ will be called a <u>design consistent regression estimator</u>.

2.3. <u>Conditional Variance</u>

We will call $\text{var}_\varepsilon(\hat{y}-y_N)$ the <u>conditional variance</u> of $\bar{y}$. Given (3) it is

$$\text{var}_\varepsilon(\hat{y}-\bar{y}_N)=$$
$$[N^{-2}(\mathbf{t}'\mathbf{V}_n\mathbf{t}-2\mathbf{t}'\mathbf{V}_n\mathbf{1}_n-2\mathbf{t}'\mathbf{V}_*\mathbf{1}_{N-n}+\mathbf{1}_N'\mathbf{V}_N\mathbf{1}_N)]+$$
$$[2N^{-2}(\mathbf{1}_N'\mathbf{X}_N-\mathbf{t}'\mathbf{X}_n)\mathbf{C}\{\mathbf{V}_n(\mathbf{t}-\mathbf{1}_n)-\mathbf{V}_*\mathbf{1}_{N-n}\}]+$$
$$[N^{-2}(\mathbf{1}_N'\mathbf{X}_N-\mathbf{t}'\mathbf{X}_n)\mathbf{CV}_n\mathbf{C}'(\mathbf{X}_N'\mathbf{1}_N-\mathbf{X}_n'\mathbf{t})] \qquad (4)$$

where $$\mathbf{V}_N=E(\boldsymbol{\varepsilon}_N\boldsymbol{\varepsilon}_N') = \begin{bmatrix} \mathbf{V}_n & \mathbf{V}_* \\ \mathbf{V}_*' & \boldsymbol{\Lambda} \end{bmatrix}.$$

When $\mathbf{C}=\mathbf{C}_W$ and both $\mathbf{W}$ and $\mathbf{V}_N$ are diagonal with bounded elements, the first bracketed term in (4) is of order 1/n, the second $O_p(n^{-3/2})$ (p again denotes selection probability), and the third $O_p(n^{-2})$. Note that while we our focusing on a model-based property of $\hat{y}$, we nonetheless employ an asymptotic consequence of $\hat{y}$ being design consistent; namely, that each element of $(\mathbf{t}'\mathbf{X}_n-\mathbf{1}_N'\mathbf{X}_N)/N$ is $O_p(n^{-1/2})$.

The first term of (4) dominates asymptotically and is independent of $\mathbf{C}$ and thus $\mathbf{W}$. As Wright (1983) noted, every bounded diagonal choice for $\mathbf{W}$ in $\mathbf{C}_W$ results in a design consistent regression estimator with the same asymptotic model variance. Tam (1986) showed that for general $\mathbf{V}_N$ an optimal $\mathbf{W}$ will exist and equal $\mathbf{V}_n$ only when $\mathbf{V}_n(\mathbf{t}-\mathbf{1}_n)-\mathbf{V}_*\mathbf{1}_{N-n}=\mathbf{X}_n\mathbf{g}$ for some k-vector $\mathbf{g}$.

In many single stage surveys, $\mathbf{V}_N$ can be assumed to be diagonal with apparent correlations across units modeled explicitly using dummy variables. Until noted otherwise in the final section, we will restrict our attention to single stage surveys and diagonal $\mathbf{V}_N$.

Formally, the restrictions on $\mathbf{V}_N$ and $\mathbf{W}$ are

$$0<M_3<v_i<M_4 \qquad (5a)$$
$$0<M_5<w_i<M_6, \qquad (5b)$$

where $\mathbf{V}_N=\text{diag}\{v_1, \ldots, v_n\}$, and $\mathbf{W}=\text{diag}\{w_1, \ldots, w_n\}$.

3. VARIANCE ESTIMATION

The general approach to variance estimation taken here is to begin with a design consistent estimator of the design mean squared error of $\hat{y}$ (assuming one exists). This mean squared error estimator, r_{YG}, is then multiplied by a factor that removes the model bias from r_{YG} as an estimator of the conditional variance of $\hat{y}$, yet is asymptotically unity. As a result, the new variance/mean squared error estimator is simultaneously a design consistent estimator of the design mean squared error of $\hat{y}$ and a model unbiased estimator of the conditional variance of $\hat{y}$.

3.1. Design Mean Squared Error

If $\mathbf{b}$ in (1) is set equal to zero, then $\hat{y}$ becomes the design unbiased Horvitz-Thompson (1952) estimator, $\hat{y}_{HT}$. The design variance of $\hat{y}_{HT}$ can be expressed as
$$var_p(\hat{y}_{HT})=N^{-2}\sum_{i<j}^{N}(p_ip_j-p_{ij})(y_i/p_i-y_j/p_j)^2.$$

When all the p_{ij} are greater than zero, a design unbiased estimator of $var_p(\hat{y}_{HT})$ is the Yates-Grundy (1953) estimator (see also Sen 1953):
$$r_{YG}=N^{-2}\sum_{i<j}^{n}[(p_ip_j-p_{ij})/p_{ij}](y_i/p_i-y_j/p_j)^2 \quad (6)$$
This estimator is itself a Horvitz-Thompson estimator based on an $\binom{n}{2}$ sample of ij pairs. Consequently, sufficient conditions for r_{YG} to be $O(n^{-2})$, and thus design consistent, are (in addition to $|y_i|$ being bounded, which is part of (2a))
$$|p_ip_j-p_{ij}|/p_i^2<M_7 \quad (2d)$$
$$N^2p_{ij}/n^2>M_8>0 \quad (2e)$$
and $\sum^N\sum^N\sum^N\sum^N h_{ijkg}/n^2<M_9$ (2f)

where
$$h_{ijkg}=p_{ij}p_{kg}-p_{ijkg} \text{ when } p_{ij}p_{kg}-p_{ijkg} \text{ is positive}$$
$$=0 \quad \text{otherwise,}$$
and p_{ijkg} is the joint probability of selecting units i, j, k, and g for the sample.

Returning to (1), suppose $\mathbf{b}=\hat{\boldsymbol{\beta}}=\mathbf{C}_W\mathbf{y}_n$. Following Fuller (1975), it is not unreasonable to assume that
$$\text{plim}_p\mathbf{b}=\mathbf{b}_* \quad (2g)$$
for some $\mathbf{b}_*$, and
$$(\mathbf{b}-\mathbf{b}_*)'(\mathbf{b}-\mathbf{b}_*)=O(n^{-1}). \quad (2h)$$
If the model in (3) holds, $\mathbf{b}_*=\boldsymbol{\beta}$. The model need not hold for $\mathbf{b}_*$ to exist however.

Let $u_i=y_i-\mathbf{x}_i\mathbf{b}_*$. The difference $\hat{y}-\bar{y}_N$ can be re-expressed as
$$\hat{y}-\bar{y}_N=N^{-1}(\mathbf{t}'\mathbf{u}_n-\mathbf{1}_N'\mathbf{u}_N) + N^{-1}(\mathbf{t}'\mathbf{X}_n-\mathbf{1}_N'\mathbf{X}_N)(\mathbf{b}_*-\mathbf{b}).$$
Consequently, the design mean squared error of $\hat{y}$ is equal to
$$E_p[(\hat{y}-\bar{y}_N)^2]=var_p(\mathbf{t}'\mathbf{u}_n/N) + O(n^{-3/2}).$$
If $\mathbf{u}_n$ were known, the design variance of $\mathbf{t}'\mathbf{u}_n/N$ could be consistently estimated with the Yates-Grundy estimator with the u_i replacing the y_i in (6), which in turn would be a design-consistent estimator of $MSE_p(\hat{y})$.

Unfortunately, the u_i are not known. Let $e_i=u_i-\mathbf{x}_i(\mathbf{b}-\mathbf{b}_*)$, so that e_i is $y_i-\mathbf{x}_i'\mathbf{b}$. It is now a simple matter to show that
$$r_{YG}=N^{-2}\sum_{i<j}^{n}[(p_ip_j-p_{ij})/p_{ij}](e_i/p_i-e_j/p_j)^2$$
is a design consistent estimator of $MSE_p(\hat{y})$ under the restrictions imposed on the sampling design and population by the various parts of equation (2).

3.2. Conditional Variance Estimation

Given the model in (3) and a known variance matrix, $\mathbf{V}_N$, satisfying (5a), the conditional variance of $\hat{y}$ is expressed in equation (4). The Yates-Grundy mean squared error estimator of $\hat{y}$, r_{YG}, has a model expectation of
$$E_\xi(r_{YG})=N^{-2}\sum_{i<j}^{n}[(p_ip_j-p_{ij})/p_{ij}]\mathbf{d}_{ij}'\mathbf{T}(\mathbf{I}_n-\mathbf{X}_n\mathbf{C})\mathbf{V}(\mathbf{I}_n-\mathbf{C}'\mathbf{X}_n')\mathbf{T}\mathbf{d}_{ij} \quad (7)$$
where $\mathbf{T}$ is a n x n diagonal matrix with t_i as its ith diagonal element,
and $\mathbf{d}_{ij}$ is an n-vector with 1 as its ith element, -1 as its jth element, and 0's elsewhere.

Consider this variance estimator:
$$r_V=[var_\xi(\hat{y}-\bar{y}_N)/E_\xi(r_{YG})]r_{YG}. \quad (8)$$
It is a model unbiased estimator of the conditional variance of $\hat{y}$. It is also a design consistent estimator of the design mean squared error of $\hat{y}$, because, as we will see shortly, the ratio adjustment factor $R_V=var_\xi(\hat{y}-\bar{y}_N)/E_\xi(r_{YG})$ is asymptotically unity even when the model in equation (3) fails. This is true not only when $\mathbf{V}_N$ is misspecified (as in Royall and Cumberland 1978), but also when $E_\xi(\mathbf{y}_N)$ does not equal $\mathbf{X}_N\boldsymbol{\beta}$!

Let the numerator of R_V be A and the denominator be B. When (3) is not true, A is simply the right hand side of (4) and B the right hand side of (7), where $\mathbf{V}_N$ is a known diagonal matrix with no particular meaning.

Now let $\boldsymbol{\eta}_N$ be a random N-vector with mean 0 and variance $\mathbf{V}_N$ Clearly, $A=var(\mathbf{t}'\boldsymbol{\eta}_n/N-\mathbf{1}_N'\boldsymbol{\eta}_N/N)+O_p(n^{-3/2})$, while $B=E_\xi E_p[(\mathbf{t}'\boldsymbol{\eta}_n/N-\mathbf{1}_N'\boldsymbol{\eta}_N/N)+O(n^{-2})$. Thus $A=B+O_p(n^{-3/2})$, and $R_V=1+O_p(n^{-1/2})$. QED.

When the model in (3) does holds and $\mathbf{V}_N$ is known up to a constant, equation (8) can be used to construct a model unbiased estimator of the conditional variance of $\hat{y}$. In many practical applications, however, a statistician will have some doubt about his (her) choice for $\mathbf{V}_N$. Consequently, we henceforth draw a distinction between $\mathbf{F}_N$, one's choice for $\mathbf{V}_N$, and the true $\mathbf{V}_N$ (supposing, of course, that (3) holds and $\mathbf{V}_N$ exists).

It is easy to see that under equations (2) and (5), any diagonal choice for $\mathbf{F}_N$ yields an estimator of the conditional variance with a relative model bias no greater than order $n^{-1/2}$. In fact, r_{YG}, as an estimator of the conditional variance has a relative model bias of the same order.

Sample sizes are not arbitrarily large, however, so the asymptotic model unbiasedness of r_{YG} should not deter us from seeking an even less biased conditional variance estimator. A reasonably chosen $\mathbf{F}_N$ will surely do better than the implied choice - which may not even exist - that results in $R_F=1$ and $r_F=r_{YG}$. Moreover, as we shall see, in certain circumstances it may not be necessary to choose values for the v_i, while in others it may be possible estimate the v_i from the sample.

4. SOME SPECIAL CASES

The Yates-Grundy mean squared error estimator collapses to a much simpler form under many sampling designs in common practice. In this section, we focus on a few of them.

In finite population sampling theory, many results simplify for large populations. We will say that a population is <u>relatively large</u> (compared to the sample) when $1/N$ is $O(n^{-3/2})$.

When the model bias of a conditional variance estimator, r, is $O_p(n^{-2})$, rather than $O_p(n^{-3/2})$ like r_{YG}, we will say it is <u>almost model unbiased</u>. Similarly, when $r'-r''$ is $O_p(n^{-2})$ we will say that r' and r'' are <u>almost equal</u>.

4.1. <u>Simple Random Sampling</u>

Under simple random sampling (srs), $p_i=n/N$ and and $p_{ij}=n(n-1)/[N(N-1)]$ for $i\neq j$. The conditional variance of $\hat{y}$ in (1) given (2), (3), and (5) is

$$\begin{aligned}\mathrm{var}_{\varepsilon}(\hat{y}-\bar{y}_N)&=(1-n/N)\bar{v}_n/n+(\bar{v}_N-\bar{v}_n)/N + \\ &2(1-n/N)(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_n)(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{v}_n + \\ &(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_n)(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{V}_n \\ &\mathbf{W}^{-1}\mathbf{X}_n(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_n)',\end{aligned} \tag{9}$$

where $\bar{v}_m=\mathbf{1}_m'\mathbf{v}_m/m$ and $\bar{\mathbf{x}}_m=m^{-1}(\mathbf{1}_m'\mathbf{X}_m)$.

The Yates-Grundy mean squared error estimator has this simplified expression:

$$r_{YG}=(1-n/N)/(n[n-1])\Sigma^n(e_i-\bar{e}_n)^2,$$

where $\bar{e}_n=(\mathbf{I}_n-\mathbf{X}_n(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1})\mathbf{y}_n$.

The model expectation of r_{YG} is

$$\begin{aligned}E_{\varepsilon}(r_{YG})&=(1-n/N)\bar{v}_n/n + (1-n/N)/(n[n-1]) \\ &\mathrm{tr}[-2(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{V}_n(\mathbf{I}_n-n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}_n \\ &+(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{V}_n\mathbf{W}^{-1}\mathbf{X}_n \\ &(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{W}_n)^{-1}\mathbf{X}_n'(\mathbf{I}_n-n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}_n]\end{aligned} \tag{10}$$

Suppose we believe that $v_i \propto f_i$. The conditional variance estimator, r_F, is then $R_F r_{YG}$, where R_F is $\mathrm{var}_{\varepsilon}(\hat{y})$ in (9) divided by $E_{\varepsilon}(r_{YG})$ in (10) with the v_i in each equation replaced by the approprepriate f_i.

In the very simple case where $\mathbf{x}_i$ contains a lone element x_i and $w_i=x_i$, $\hat{y}$ collapses into the ratio estimator, $\hat{y}_R=\bar{x}_N\bar{y}_n/\bar{x}_n$. The conditional variance estimator, r_F, is a complicated expression that corresponds exactly to the estimator in Royall and Eberhardt (1975) when $f_i \propto x_i$.

When the population is relatively large,

$$r_F=[1+2(\bar{x}_N-\bar{x}_n)/\bar{x}_n]r_{YG} + O_p(n^{-2}). \tag{11}$$

It is not difficult to see that the dominant part of (11) is almost equal to this familiar form $r_2=(\bar{x}_N/\bar{x}_n)^2 r_{YG}$ (e.g., see Cochran, 1977, p. 155). Thus, r_2 is almost model unbiased no matter what $\mathbf{V}_N$ truly is. On the other hand, no $\mathbf{V}_N$ will produce a r_V almost equal to r_{YG}.

Wu and Deng (1983) empirically studied variance estimators of the form $r_g=(\bar{x}_N/\bar{x}_n)^g r_{YG}$; g=0, 1, 2. They found that as an estimator of the unconditional, design mean squared error of $\bar{y}_N$, different choices for g fit better for different populations. As an estimator of the squared error of $\hat{y}_R$ conditioned on a realized $\bar{x}_n$, however, r_2 fit best in every population. This latter result is consistent with our theory.

For more general $\mathbf{x}_i$, the choice of $\mathbf{F}_N$ may matter. Royall and Cumberland (1978) suggest using the nearly unbiased $e_{iA}^2=e_i^2/(1-\mathbf{x}_i(\mathbf{X}_n'\mathbf{W}^{-1}\mathbf{X}_n)^{-1}\mathbf{x}_i'w_i^{-1})$ as estimators of the v_i, i=1, ..., n. This approach is reasonable when the population is relatively large so that $(\bar{v}_n-\bar{v}_N)/N$ in (9) can (almost) be ignored. (N.B. Since Royall and Cumberland did not invoke the asymptotic proporties of randomization, they were forced to assume that $1/N$ was even less than $O(n^{-3/2})$.)

Alternatively, when N is not relatively large, we we may have reason to believe that

$$\varepsilon_i^2=\mathbf{z}_i\boldsymbol{\gamma}+\theta_i,$$

where the θ_i are random variables with mean zero, $\mathbf{z}_i$ is a known row vector, and $\boldsymbol{\gamma}$ is an unknown column vector. If this is the case, then regressing the e_{iA}^2 on the $\mathbf{z}_i$ seems a reasonable procedure for estimating $\boldsymbol{\gamma}$ and through it the f_i; i.e., $f_i=\hat{v}_i=\mathbf{z}_i\hat{\boldsymbol{\gamma}}$.

4.2. <u>Hartley-Rao Sampling</u>

Suppose that (2), (3), and (5) hold and $\mathbf{x}_i=x_i$. In addition, assume that our best guess before sampling is that $\mathbf{V}_N=\mathbf{F}_N$ (we may have another guess after samplng).

Let

$$\hat{y}=\Sigma^n y_i/(Np_i) + (\bar{x}_N-\Sigma^n x_i/[Np_i])\hat{\beta}_W,$$

where $\hat{\beta}_W=\Sigma^n y_i x_i w_i^{-1}/\Sigma^n x_i^2 w_i^{-1}$.

The most asymptotically efficient estimation strategy involving an estimator like $\hat{y}$ sets

$$p_i=n(f_i)^{1/2}/\Sigma^N(f_j)^{1/2} \tag{12}$$

(Brewer 1963). Hartley and Rao (1962) discuss and analyze a method of sampling that yields (12) - systematic probability proportional to size sampling from a randomly ordered list. They also propose a useful approximation of the Yates-Grundy mean squared error estimator for relatively large populations.

When $w_i=x_i p_i$, $\hat{y}$ collapses to the (weighted) ratio:

$$\hat{y}=\bar{x}_N[\Sigma^n(y_i/p_i)/\Sigma^n(x_i/p_i)].$$

Other suggested values for w_i are $w_i=f_i$ (Little 1983) and $w_i=f_i p_i$ (Sarndal 1982).

When $p_i \propto x_i$ we have the standard Horvitz-Thompson estimator. Since $\Sigma^n x_i/(p_i N) \equiv x_N$ the choice of $\mathbf{W}$ becomes irrelevant in this special case.

For a relatively large population, the conditional variance of $\hat{y}$ is (from (4))

$$\begin{aligned}\mathrm{var}_{\varepsilon}(\hat{y}-\bar{y}_N)&=N^{-2}(\Sigma^n v_i/p_i^2-2\Sigma^n v_i/p_i+\Sigma^N v_i) + \\ &2N^{-1}(\bar{x}_N-\Sigma^n x_i/(p_i N)) \\ &\Sigma^n v_i x_i w_i^{-1} p_i^{-1}/\Sigma^n x_i^2 w_i^{-1} + O_p(n^{-2}).\end{aligned}$$

Call the dominant part of this expression A_V.

Using the Hartley-Rao relatively large population approximation, the Yates-Grundy design mean squared error of $\hat{y}$ is

$$r_{YG}=[2N^2(n-1)]^{-1}\sum_{i\neq j}^{n}(1-p_i-p_j+\sum^N p_k^2/n)(e_i/p_i-e_j/p_j)^2,$$

where $e_i=y_i-x_i\hat{\beta}_W$, and its model expectation is

$$E_\varepsilon(r_{YG})=N^{-2}[\sum^n v_i/p_i^2(1-\sum^n p_j/n+\sum^N p_k^2/n)-\sum^n v_i/p_i]+O(n^{-2})..$$

Call the dominant part of this expression B_V.

Now for relatively large populations,

$$r_V^*=(A_V/B_V)r_{YG}.$$

is an estimator of the conditional variance of $\hat{y}$ that is almost model unbiased when $\mathbf{V}_N$ is known and a design consistent estimator of the design mean squared error of $\hat{y}$ even when the model in (3) fails.

What if the model holds but $\mathbf{V}_N$ is unknown? When $\hat{y}$ is in the form of the ratio, and the population is relatively large, it is not difficult to see that

$$r_2=(\bar{x}_N/\hat{x}_n)^2 r_{YG} \tag{13}$$

where $\hat{x}_n=\sum^n x_i/(p_iN)$, is almost equal to r_V^*. (We are using the fact that $\sum^n v_i/(p_iN)-\bar{v}_N$ and $(N/n^2)(\sum^n p_i-\sum^N p_i^2)$ are $O_p(n^{-1/2})$.)

This suggests that although all choices for **W** are asymptotically identical as far as the model efficiency and design consistency of $\hat{y}$ are concerned, when it comes to estimating the variance of $\hat{y}$, $w_i=x_ip_i$ produces a conditional variance estimator with an attractive robustness when the population is relatively large.

For the Horvitz-Thompson estimator $\hat{x}_n\equiv\bar{x}_N$. Consequently, as noted by Cumberland and Royall (1981), r_{YG} is an almost model unbiased estimator of the conditional variance of y when the population is relatively large.

5. POSSIBLE EXTENSTIONS

The analysis so far has been limited in a number of ways. Attention has been focused on estimating means, on linear regression estimators that are model unbiased, and on certain single stage, fixed size sampling designs. Extensions to other population parameters and other design consistent estimation strategies are possible. Much of the groundwork has been broken here. (The bounds on $|y_i|$ and $|x_{ij}|$ can also be weakened by strengthening the restrictions of the sampling designs; see lemma 1 of Isaki and Fuller (1982) for an indication of how this may be done.)

With some care it is possible to combine multistage sampling and design consistent regression estimation. Although the theoretical work in the text was confined to diagonal $\mathbf{V}_N$, it appears possible to develop the analysis for any positive definite $\mathbf{V}_m$ (m=n or N) with an order m number of non-zero elements. (The key is the asymptotic property of the ratio adjustment factor in (8) given a possibly misspecified $\mathbf{V}_N$.) This condition is often satisfied by populations undergoing multistage sampling, where only units within the same sampling cluster are assumed to be correlated.

In many multistage and other sampling designs, the sample size is not fixed so the Yates-Grundy design mean squared error estimator is invalid. Where an alternative design mean error estimator exists and is itself design consistent, the application of the basic ratio adjustment technique for simultenously producing a(n) (almost) model unbiased conditional variance and a design consistent design mean squared error estimator should still apply.

It would be incorrect to infer from the text that model-based conditional variance estimators are unavailable for sampling designs that fail to satisfy equation (2). Quite the opposite. It is design-based mean squared error estimators that do not exist for such plans. Any attempt to calcuate variance must then be model-based. This is a point not often enough recognized by survey statisticians.

REFERENCES

Brewer, K. R. W. (1963), "Ratio Estimation and Finite Populations: Some Results Deducible 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 Assocoation, 74, 911-915.

Cochran, W. G. (1977), Sampling Techniques, 3rd edition, New York: John Wiley and Sons.

Cumberland, W. G. and Royall R. M. (1981), "Prediction Models and Unequal Probability Sampling," Journal of the Royal Statistical Society, Ser. B., 43, 353-367.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," Sankhya, Ser. C, 37, 117-132.

_____ (1981), "Comment," Journal of the American Statistical Association, 76, 78-80.

_____ and Isaki, C. T. (1981), "Survey Design Under Superpopulation Models," in Current Topics in Survey Sampling, D. Krewski, J. N. K. Rao, and R. Platek eds., New York: Academic Press, 199-225.

Hartley, H. O. and Rao, J. N. K. (1962), "Sampling With Unequal Probabilities and Without Replacement," Annals of Mathematical Statistics, 33, 350-374.

Hidiroglou, M. A. and Sarndal C. E. (1986), "Conditional Inference for Small-Area Estimation," American Statistical Association Proceedings of the Section on Survey Research Methods, 147-155.
Holt, D. and Smith, T. M. F. (1979), "Post Stratification," Journal of the Royal Statistical Society, Ser. A, 142, 33-46.
Horvitz D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe, Journal of the American Statistical Association, 47, 663-685.
Isaki, C. T. and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," Journal of the American Statistical Association, 77, 89-96.
Kott, P. S. (1986), "Some Asymptotic Results for the Systematic and Stratified Sampling of a Finite Population," Biometrika, 73, 485-491.
Little, R. J. (1983), "Estimating a Finite Population Mean from Unequal Probability Samples," Journal of the American Statistical Association, 78, 596-604.
Oh, H. L. and Scheuren, F. J. (1983), "Weighting Adjustment for Unit Nonresponse," in Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, W. G. Madow, I. Olkin, and D. B. Rubin eds., 143-83.
Rao, J. N. K. (1985), "Conditional Inference in Survey Sampling," Survey Methodology, 11, 15-31.
Robinson P. M. and Sarndal C. E. (1983), "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling, Sankhya, Ser. B, 45, 240-248.
Royall, R. M. (1970), "On Finite Population Sampling Theory under Certain Linear Regression Models," Biometrika, 57, 377-87.
_____ (1976), "Current Advances in Sampling Theory: Implications for Human Observational Studies," American Journal of Epidemiology, 104, 463-474.
_____ (1983), "Comment," Journal of the Amercian Statistical Association, 78, 794-796.
_____ (1986), "The Prediction Approach to Robust Variance Estimation in Two -Stage Cluster Sampling," Journal of the American Statistical Association, 81, 119-123.
_____ and Cumberland W. G. (1978), "Variance Estimation in Finite Population Sampling," Journal of the American Statistical Association, 73, 351-358.
__________ (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, 66-77.
_____ and Eberhardt, K. R. (1975), "Variance Estimates for the Ratio Estimator," Sankhya, Ser. C, 37, 43-52.
Sarndal, C. E. (1980), "On -Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling," Biometrika, 67, 639-650.
_____ (1982), "Implications of Survey Design for Generalized Estimation of Linear Functions," Journal of Statistical Planning and Inference, 7, 155-170.
Sen, A. R. (1953), "On the Estimate of Variance in Sampling with Varying Probabilities," Journal of the Indian Society of Agricultural Statistics, 5, 119-127.
Tam, S. M., (1986), Characterization of Best Model-Based Predictors in Survey Sampling," Biometrika, 73, 232-235.
Valliant R. (1987), "Some Prediction Properties of Balanced Half-Sample Variance Estimators in Single-Stage Surveys," Journal of the Royal Statistical Society, Ser B, forthcoming.
Wright, R. L. (1983), "Finite Population Sampling with Multivariate Information," Journal of the American Statistical Association, 78, 879-884.
Wu , C. F. J. (1982), "Estimation of Variance of the Ratio Estimator," Biometrika, 69, 183-189.
_____ (1985), Variance Estimation for the Combined Ratio and Combined Regression Estimators," Journal of the Royal Statistical Society, Ser. B, 47, 147-154.
_____ and Deng L. Y. (1983), "Estimation of Variance of the Ratio Estimator: An Empirical Study," in Scientific Inference, Data Analysis, and Robustness, L. Box and Wu eds., New York: Academic Press, 245-277.
Yates F. and Grundy, P. M. (1953), "Selection Without Replacement from Within Strata with Probability Proportional to Size," Journal of the Royal Statistical Society, Ser. B., 15, 253-261.