VARIANCE ESTIMATION USING PSEUDOSTRATA FOR A LIST-SUPPLEMENTED AREA PROBABILITY SAMPLE

Miriam L. Goldberg and Paul M. Gargiullo, Energy Information Administration
Miriam L. Goldberg, EI-652, U.S Department of Energy, Washington, DC 20585

KEY WORDS: multiple frame sampling, pseudo-strata, jackknife, complex survey

Introduction

Supplementing an area sample with a sample from a list frame of important units can be an effective way to improve the efficiency of a multistage sample. This technique is useful for sampling from a population where the units of greatest interest are relatively rare. To form unbiased estimators, a theoretically straightforward approach would be to calculate joint probabilities of selection for sample units in the overlap between the area and list portions of the sample. However, this can be very difficult to accomplish for multistage area samples, since listing of basic units is restricted to selected higher stage units. Approaches that have been used in some surveys adjust certain sample weights to account for the overlap between the area and list frames; these adjustments produce unbiased estimates of population totals without the difficulties of calculating joint selection probabilities.

This paper derives the components of variance of the estimates produced using one type of overlap adjustment procedure, and discusses the construction of pseudostrata to capture these variance components. The methods are described in reference to the Department of Energy's Nonresidential Buildings Energy Consumption Survey (NBECS).

The 1986 NBECS was based on a four-stage area probability sample, where individual buildings were the ultimate sample units. For the first stage of sampling, 129 strata were formed by grouping together similar primary sampling units (PSU's). Each of the 1,799 PSU's was composed of a city and surrounding counties, or of rural counties only. Thirty-two of the PSU's were highly populated, and were not grouped with other PSU's to form strata. These 32 PSU's were designated as certainty PSU's and were taken into the sample with probabilities of 1. For the remaining strata, containing grouped noncertainty PSU's, one PSU was selected from each stratum, yielding a total first-stage sample of 129 PSU's. Within each selected PSU, further sampling stages selected ZIP code groups, then area segments, and finally buildings.

Very large buildings, although relatively rare in the population, account for a high proportion of total energy consumption. To ensure adequate coverage of large buildings and of others that were significant energy users, the area sample within each PSU was supplemented by a sample from lists of large and "special" buildings such as hospitals and schools. For both area and list samples, the overall selection probabilities for individual buildings were set to be proportional to building size, which is correlated with energy consumption. The supplementary sample from the lists of large and special buildings thus improves the overall efficiency of the sample design. This is true even though these lists have undercoverage and do not comprise a complete within-PSU sampling frame.

Constructing the NBECS Linear Estimator

Supplementing an area probability sample with a sample from list frames requires special treatment for the frame overlap, to avoid double counting and produce unbiased survey estimates. The most theoretically straightforward way to create unbiased linear estimators is to compute (with great difficulty) the joint probabilities of selection for those sample units that can be selected into either the area sample or the list sample.

Another way to handle the overlap between the area sample and the list frame is simply to delete any cases from the area sample found to be on the list frame. This is the "screening" estimator approach used by the National Agricultural Statistics Service (NASS) in its surveys of crops and liverstock (Bosecker and Ford 1976). For the NBECS, though, simply eliminating the intersection cases (unless also sampled from the lists) would be an inefficient use of resources. The information required to determine that an area-sample building is on the list frame is typically not obtained until the time of interview. At that point, the incremental cost of completing the interview and obtaining usable information is small compared to the cost already incurred of getting to the building.

Composite estimators offer an alternative approach, and are currently in use by NASS (Hartley 1962, and Bosecker and Ford 1976). Two independent estimates of the list-frame universe are obtained, one from the area sample, using area-sample weights, and one from the list sample, using list-based weights. An overall estimate of the list-frame universe is then constructed as the linear combination of these two separate estimates, with weights in inverse proportion to the variances of the two estimates. In principle, this would be the most efficient way to combine the information from the intersection and list samples. However, this approach is impractical for the NBECS, which is a multi-purpose survey used to produce extensive tabulations of population aggregates. The increased computational complexity is unlikely to be justified by the improvement in resulting estimates.

A simpler unbiased procedure in current use for NBECS is formulated in terms of weighting rules as follows:

1) Area-sampled selections that are not included in any of the list frames are given the appropriate area-sample weights. These sample selections thus provide an estimate of only that portion of the total population that does not appear on any list frames.

2) All area-sampled buildings that are included on a list frame are assigned a within-PSU weight of 1. That is, they represent only themselves within the PSU.

3) List-sampled selections that are not also taken into the area sample are assigned the appropriate list-sample weights. These buildings provide an estimate of that portion of the list frames that did not also come into the area sample.

The resulting estimator x' is the sum of three components,

$$x' = A + I + N,$$

where A is a linear combination of attributes of sample units from the area sample, excluding buildings appearing on any list frame; I is a linear combination of units from the intersection sample, consisting of buildings on the list frame that were also selected into the area sample; and N is a linear combination of units from the nonintersection list sample, consisting of buildings on the list frame and sampled only in the list sample.

This estimator was originally proposed by Jack Ogus, a consultant to the 1979 NBECS contractor.

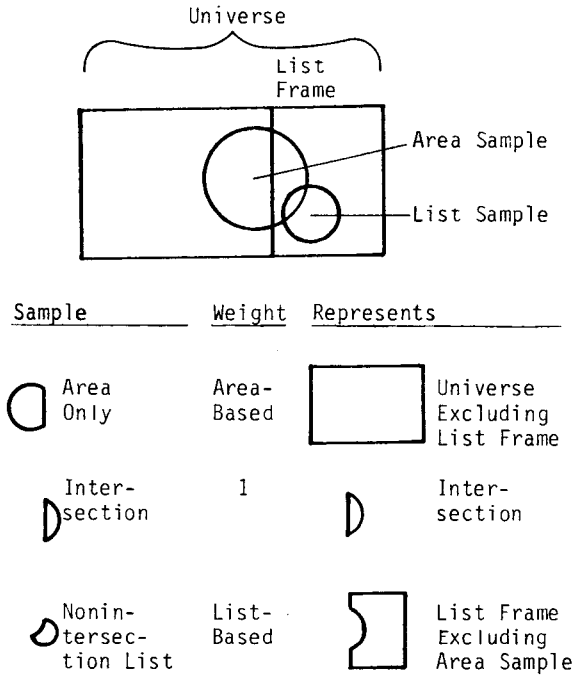
It can be shown (Chu, 1987) that for any PSU-level aggregate X, the linear estimator (x') resulting from this weighting procedure is unbiased.

As illustrated in Figure 1, the components I and N together estimate the list-frame aggregate, while the component A estimates the aggregate for the population excluded from the list frame. The present paper explores the contributions of each of the three components A, I, and N to the variance of x', and the relationships among these contributions.

Estimating the Variance of the Linear Estimator

To estimate variances of survey statistics in NBECS, the first-stage sampling strata were collapsed to form pseudostrata, and the sampled PSU's paired in this way were used in a jackknife replication technique (McCarthy, 1966 and 1969). One goal of the present work is to determine the appropriate treatment of the intersection buildings in constructing variance-estimation pairs (or pseudostrata). It has been the convention in the past to treat these buildings as certainty units (since they have weights of 1), and include them in all replicates, rather than placing them into variance-estimation pairs so that they would be left out of some replicates and included in others. This convention implies that these intersection buildings have zero contribution to the variance of x'. This is actually not the case. The main conclusion of the developments described below is that the convention results in a slight overestimate of variance, but is adequate for present applications. Alternative methods have been considered, and may be used in future surveys.

Figure 1. NBECS Overlap Adjustment Scheme



Before proceeding to the main results, some general principles of constructing variance-estimation pairs, and the application of those principles to the NBECS sample, are reviewed.

General Principles for Constructing Variance Estimation Pairs

The variance and covariance relationships developed here are for linear estimators, which are linear combinations of sample values. Linear estimators include weighted sums of sample-building attributes, provided the weights are nonrandom. Variance estimation pairs are constructed to provide unbiased estimates of variance for such statistics.

Within each stratum, this construction pairs together two sampling units at the earliest stage of sampling, in such a way that the two members represent two independent samples, each of which incorporates all subsequent stages of random selection. Restricting calculations to the initial stage of the sample is called the "Ultimate Cluster" technique (Wolter, 1985). In practice, ultimate clusters often refer to noncertainty PSU's, and to SSU's within certainty PSU's. The difference between estimates based on these two ultimate clusters then gives an estimate, with one degree of freedom, of the variance of the component of the statistic contributed by the initial sampling stage.

If only one first-stage unit was sampled from each of the original first-stage strata, the variance-estimation pairs must be constructed by collapsing the original strata into pseudo-strata. Only similar strata are collapsed together, in order to minimize bias in the variance estimates due to between-strata components. Stratum similarity must be judged using only pre-sampling information, to avoid further bias.

Sets of ultimate clusters that were drawn independently of each other can be split into pairs separately to represent independent variance components. But sets whose contributions to the aggregate estimate have nonzero covariance need to be combined into a single pair if this covariance is to be correctly reflected in the variance estimate.

More specifically, if the variances of B and C, respectively are estimated by

$$(B1 - B2)^2 \quad \text{and} \quad (C1 - C2)^2,$$

then the combined pair

$$((B1 + C1) - (B2 + C2))^2 = (B1 - B2)^2$$

$$+ \ (C1 - C2)^2 + 2\,(B1 - B2)(C1 - C2)$$

represents the variance of the sum B + C, with the cross-product of differences estimating the covariance of B and C. If B and C are kept as separate pairs, this covariance term will not be represented in the resulting variance estimate, resulting in a bias if the covariance is nonzero.

If, on the other hand, the covariance is zero, then keeping B and C as separate pairs gives a more accurate estimate of the total variance. If the pairs are merged, the cross-product term will have zero expectation, but will contribute to the variance of the variance estimate. Pseudostrata were formed for the 1986 NBECS in such a way as to minimize the additional variance of the variance estimate that would result from cross-product terms with zero expectation.

Application of General Principles to the NBECS Sample

For the NBECS noncertainty PSU's, the first stage of sampling is the selection of PSU's themselves. Preliminary pseudostrata were formed in this group by pairing one PSU (one ultimate cluster) against another. All stages of sampling within each PSU were conducted independently of the sampling in other PSU's. Thus, the difference between PSU's correctly reflects the variance of the entire selection and estimation process involving noncertainty PSU's. This process included the selection of some intersection cases, and the setting of their within-PSU weights to one. Thus, it is appropriate to keep these intersection cases with the rest of their PSU for the pairing; this has been done for the NBECS.

Construction of variance-estimation pairs for certainty PSU's is the subject of the remainder of this paper. For the NBECS certainty PSU's, the first sampling stage is the secondary sampling unit or SSU. Each SSU selection for the sample is independent of any other SSU selection. A single segment is selected within each of the independently selected SSU's.

Variances and Covariances of the Linear Estimator Components

The derivations that follow are for segment-level variance and covariance relationships. All these results hold also for the PSU-level variances and covariances, which, under independence assumptions, are the corresponding sums of the segment-level terms.

Lemma 1 establishes the validity of partitioning the total variance of an aggregate estimate into a component due to the area-only sample and a component due to the list-frame buildings. Formulas for these separate variance components have been derived, in terms of selection probabilities at different stages (Goldberg, 1988). These formulas offer some insight into the components of variance, but give no direct guidance for variance-pair construction. Lemmas 2 and 3 and their corollaries show relationships among the components of the list-frame variance. These relationships do have direct implications for the construction of variance estimation pairs. Lemma 4 summarizes the results of Lemmas 1, 2, and 3 to express the total variance of the within-PSU aggregate in terms of the variances of the components A, I, and N.

The estimator of a PSU-level population attribute X is given (Chu, 1987) as

$$x' = \sum_i x'_i$$

where

$$x'_i = \frac{x_{Ai}\, d_i}{P_{Ai}} + \sum_j X_{Lij}\, d_{ij} + \sum_j X_{Lij}\, (1 - d_{ij})\, g_{ij}/P_{Lij}$$

X = population value of an aggregate attribute within the PSU

x' = sample-based estimate of population aggregate X

X_{Ai} = value of aggregate in ith segment

x_{Ai} = sample-based estimate of aggregate X_{Ai}

X_{Lij} = value of the attribute for the jth list-frame building in the ith segment

d_i = indicator variable (0/1) for inclusion of segment i in the area sample

d_{ij} = indicator variable for inclusion of list-frame building j from segment i in the area sample

g_{ij} = indicator variable for inclusion of list-frame building j from segment i in the list sample

P_{Ai} = probability that segment i is included in the sample (i.e. probability that $d_i = 1$)

P_{Lij} = probability that list-frame building j from segment i is included in the list sample

= probability that $g_{ij} = 1$

Define also

P_{Aij} = conditional probability that list-frame building j from segment i is included in the area sample, given that segment i is included in the area sample

= conditional probability that $d_{ij} = 1$ given that $d_i = 1$

h_{ij} = indicator variable for inclusion of list-frame building j from segment i in the area sample, conditional on inclusion of segment i in the area sample.

That is, $d_{ij} = (d_i)(h_{ij})$, where d_i and h_{ij} are independent, and P_{Aij} is the probability that $h_{ij} = 1$. We assume that the random variables d_i, h_{ij}, g_{ij}, and x_{Ai} (which contains the random estimation error for segment i) are all mutually independent, across all i and j.

The estimator (x'_i) can be expressed as the sum of three pieces

$$x'_i = A_i + I_i + N_i$$

where

$$A_i = x_{Ai}\, d_i / P_{Ai}$$

$$I_i = \sum_j X_{Lij}\, d_i\, h_{ij}$$

$$N_i = \sum_j X_{Lij}\, (1 - d_i\, h_{ij})\, g_{ij} / P_{Lij}\,.$$

In this break-down, A_i represents the contribution from the area sample exclusive of the portion that intersects the list frame, I_i represents the contribution from that intersection, and N_i represents the contribution from the list frame not intersected by the area sample. The total list-frame contribution is denoted by

$$L_i = I_i + N_i.$$

As noted above, relationships are established here among the within-segment variances and covariances of the components A_i, I_i, N_i, and L_i. Under the assumption of independence among segments, all these relationships hold also for summations of these components over the segments, where the summations are denoted respectively as A, I, N, and L.

Lemma 1: $\mathrm{Cov}(A_i, L_i) = 0$.

Proof:

Since x_{Ai} is independent of all random terms in L_i,

$$\mathrm{Cov}(A_i, L_i) = E(x_{Ai}/P_{Ai})\, \mathrm{Cov}(d_i, L_i) = (X_{Ai}/P_{Ai})\, \mathrm{Cov}(d_i, L_i).$$

Combining the expressions for I_i and N_i gives

$$L_i = \sum_j X_{Lij}\, (d_i\, h_{ij}\, (1 - g_{ij}/P_{Lij}) + g_{ij}/P_{Lij}).$$

Since h_{ij} and g_{ij} are independent of d_i and of each other,

$$E(L_i \mid d_i) = \sum_j X_{Lij} \left(d_i P_{Aij} (1 - P_{Lij}/P_{Lij}) + P_{Lij}/P_{Lij} \right)$$

$$= \sum_j X_{Lij}$$

$$= E(L_i).$$

Therefore $Cov(d_i, L_i) = 0$, and

$Cov(A_i, L_i) = 0.$

Q.E.D.

Thus, the component of the linear estimate due to the area-sample buildings excluding those on the list frame is uncorrelated with the component due to all list-frame buildings, including both intersection and non-intersection buildings. That is, the estimates of the list-frame aggregate and its complementary population aggregate are uncorrelated.

Corollary 1: $Var(x'_i) = Var(A_i) + Var(L_i)$.

Lemma 2: $Cov(A_i, I_i) = X_{Ai} (1 - P_{Ai}) \sum_j X_{Lij} P_{Aij}$

Proof:

By the independence of x_{Ai}, d_i, and h_{ij},

$$Cov(A_i, I_i) = Cov(x_{Ai} d_i / P_{Ai}, \sum_j X_{Lij} d_i h_{ij})$$

$$= E(x_{Ai}/P_{Ai}) \, Var(d_i \sum_j X_{Lij} E(h_{ij})$$

$$= X_{Ai} (1 - P_{Ai}) \sum_j X_{Lij} P_{Aij} .$$

Q.E.D.

Corollary 1: $Cov(A_i, I_i) \geq 0$

(with equality only if)

$P_{Ai} = 1$, $X_{Ai} = 0$,

or $\sum_j X_{Lij} P_{Aij} = 0$).

Corollary 2: $Cov(A_i, N_i) = - Cov(A_i, I_i) \leq 0$

(with equality under the same conditions as for Corollary 1).

Thus, except in trivial conditions, the area-only component has nonzero covariance with the intersection component. As a consequence of Lemma 1, the covariance between the area-only and nonintersection components is of the same magnitude, but opposite sign.

Lemma 3: $E(N_i \mid I_i) = E(L_i) - I_i$.

Proof:

$$E(N_i \mid I_i) = E(\sum_j X_{Lij} (1 - d_i h_{ij}) \; g_{ij}/P_{Lij} \mid \sum_j X_{Lij} d_i h_{ij})$$

$$= \sum_j X_{Lij} (1 - d_i h_{ij}) \; E(g_{ij}/P_{Lij} \mid \sum_j X_{Lij} d_i h_{ij})$$

by the independence of all g_{ij} from d_i and all h_{ij}. This independence also means that the expectations of the g_{ij}'s are not dependent on any function of d_i and the h_{ij}'s. Hence,

$$E(N_i \mid I_i) = \sum_j X_{Lij} (1 - d_i h_{ij})$$

$$= \sum_j X_{Lij} - \sum_j X_{Lij} d_i h_{ij}$$

$$= E(L_i) - I_i.$$

Q.E.D.

Corollary 1:

$$E(L_i \mid I_i) = E(N_i \mid I_i) + E(I_i \mid I_i) = E(L_i).$$

Corollary 2: $Cov(L_i, I_i) = 0$.

Corollary 3:

$$Cov(N_i, I_i) = Cov(L_i, I_i) - Cov(I_i, I_i) = -Var(I_i).$$

That is, the covariance between the intersection and nonintersection list components is the negative of the variance of the intersection component.

Corollary 4: $Var(L_i) = Var(N_i) - Var(I_i)$.

Lemmas 1 through 3 and their corollaries lead directly to the main result: the total variance is equal to the sum of the area-only variance and the nonintersection variance minus the intersection variance. Formally, we have

Lemma 4:

$$Var(x'_i) = Var(A_i) + Var(N_i) - Var(I_i).$$

Proof:

$$\mathrm{Var}(x'_i) = \mathrm{Var}(A_i) + \mathrm{Var}(L_i)$$

$$= \mathrm{Var}(A_i) + \mathrm{Var}(N_i) - \mathrm{Var}(I_i).$$

Q.E.D.

Implications of the Variance-Covariance Relationships

1. The corollary to Lemma 1 means that the area-only sample can be split into pairs separately from the list sample. This has been the standard practice for NBECS.

2. Lemma 4 implies that treating intersection buildings as having zero contribution to the variance amounts to omitting a negative contribution to the variance, resulting in an overestimate of variance. Because the total number of intersection buildings is expected to be small in each sample, this overestimate is slight. This treatment of conditional certainty buildings has been used in the past for NBECS, and is in current use.

3. Corollary 1 to Lemma 2 means that including the intersection buildings with the area-only pairs according to the segments the intersection cases came from would result in a greater overstatement of variance, because this would add the positive covariance between the area-only and intersection, while omitting the negative covariance between area-only and nonintersection, and still omitting the negative covariance between intersection and nonintersection. Treating intersection cases this way was thus rejected.

4. Lemma 4 shows that an unbiased estimate of the total variance could be constructed by subtracting an unbiased estimate of the intersection variance from the sum of unbiased estimates of variance for the other two components. The estimate of intersection variance could be obtained by pairing intersection segments in the same way the area-only segments were paired, but not merging the intersection pairs with the area-only pairs. This approach is not recommended, because of two computational problems. First, the variance estimate could be negative in some cases, particularly for cells with small sample sizes. Second, it is not clear how to adapt either the mechanics or the theory of variance estimation by Balanced Repeated Replications or Jackknife to a variance estimator that requires the subtraction of a variance component.

An unbiased estimate of the total variance could also be formed, avoiding the problems just noted, by forming within each PSU a single stratum combining the intersection and nonintersection list samples. This combined stratum would correctly incorporate the covariance between the two into the total variance estimate. A procedure for implementing such an approach has been explored. The approach is not in use at this time because the list-sample pairing was constructed based on other considerations, including confidentiality.

5. Lemma 3 suggests that the current pairing of the nonintersection list buildings, treating all these buildings as independent selections, is not quite appropriate for estimating the variance of the nonintersection component. Because buildings are selected for the intersection sample in clusters in each PSU, and these clusters are ineligible for the nonintersection sample, there is some negative covariance among the selections made for the nonintersection sample. A theoretically correct, but more complicated method has been considered for estimating the variance of the nonintersection sample. Consideration of this method indicates that it would yield variance estimates negligibly different from those obtained under the current scheme.

References

Bosecker, R. R. and Ford, B. L. (1976) Multiple Frame Estimation with Stratified Overlap Domain. Tech. Report. Sample Survey Res. Branch, Res. Div., Stat. Reporting Serv., U.S. Dep't. of Agric., Washington, DC.

Chu, Adam (1987). Proof that the Assignment of Conditional Weights of 1 Will Produce Unbiased Estimates, in "Weighting Procedures for NBECS III". Westat, Inc., Rockville, MD (Unpublished).

-----(1988). Variance of NBECS III Estimator, in "Weighting Procedures for NBECS III". Westat, Inc., Rockville, MD (Unpublished).

Energy Information Administration (1986). Nonresidential Buildings Energy Consumption Survey: Commercial Buildings Consumption and Expenditures, 1983, U.S. Dep't. of Energy, Washington, D.C. DOE/EIA-0318(83).

-----(1985). Nonresidential Buildings Energy Consumption Survey: Characteristics of Commercial Buildings, 1983, U.S. Dep't of Energy, Washington, D.C. DOE/EIA-0246(83).

Goldberg, Miriam L. (1988) "Treatment of Conditional Certainty Buildings in the NBECS Variance Estimator". NBECS Technical Note No. 62, U.S. Energy Information Administration, Energy End Use Division, Washington, DC 20585.

Hartley, H.O. (1962) Multiple Frame Surveys. Proceedings of the Social Science Section of the American Statistical Association.

McCarthy, Philip J. (1969). Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique. Vital and Health Statistics. HEW Pub. No. (HSM) 73-1270, Series 2 - No. 31. Nat'l Center for Health Stat., PHS, Washington, D.C.

-----(1966). Replication: An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. HEW Pub. No. (PHS) 79-1269, Series 2 - No. 14. Nat'l Center for Health Stat., PHS, Washington, D.C.

Wolter, Kirk M. (1985) Introduction to Variance Estimaton. Springer-Verlag, New York.