UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 38

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: Mary H. Mulry *(signed 12/31/02)*
Chair, A.C.E. Revision II Assessment Subgroup

Prepared by: Mary H. Mulry,
Principal Researcher,
Statistical Research Division

Subject: A.C.E. Revision II – Study Plan for Confidence Intervals and Loss Function Analysis

1. BACKGROUND

Two methods of assessing the relative accuracy of the Census and the A.C.E. Revision II are using confidence intervals for the adjustment factors and a loss function analysis. We form the confidence intervals for the net undercount rate using estimates of net bias and variance. Since most of the data available on the quality of the original A.C.E. are being incorporated in the A.C.E. Revision II, the estimation of the net bias will use the data that were not included. In the loss function analysis, we use weighted squared error loss, which also may be described as the weighted Mean Squared Error (MSE), for levels and shares for counties and places across the nation and within state.

These methods for evaluating the accuracy of the census and an adjustment of the census have been used previously (Mulry and Spencer 1993, 2001; CAPE 1992).

2. QUESTION TO BE ANSWERED

Is the A.C.E. Revision II or the census more accurate for shares and levels for states, counties, and places?

3. METHODOLOGY

Appendix A contains a detailed description of the methodology for the construction of the confidence intervals for the undercount rate while Appendix B describes the loss function methodology.

The construction of the confidence intervals incorporate both sampling and nonsampling error. Since most of the data available on the quality of the original A.C.E. is being incorporated in the A.C.E. Revision II, the estimation of the net bias will use the data that was not included. The bias combined the error due to inconsistent reporting of variables used in poststratification (Bench 2002), the error due to using the inmovers to represent the movers in the PES-C formulation of the dual system estimator (DSE) (Keathley 2002), and the error in the identification of duplicate enumerations in the census as measured by administrative records (Bean and Bauder 2002). The estimate of the variance in A.C.E. Revision II included three error components. These are the sampling error, the error due to the choice of the missing data model (Kearney 2002), and the error due to the choice of model for correcting for duplicate enumerations (Davis 2002).

Confidence intervals that incorporate the net bias as well as the variance for the undercount rate $\hat{U}$ provide a method for comparing the relative accuracy of the census and the A.C.E. Revision II estimates. We will estimate the net bias in the census coverage correction factor for each poststratum. With the estimated bias and variance for each evaluation poststrata, we can estimate the bias $\hat{B}(\hat{U})$ and variance V in the net undercount rate $\hat{U}$ and form the 95% confidence interval for the net undercount rate by

$$(\hat{U}-\hat{B}(\hat{U})-2\hat{V}^{1/2},\hat{U}-\hat{B}(\hat{U})+2\hat{V}^{1/2}).$$

Since $\hat{U}=0$ corresponds to no adjustment of the census, one comparison of the relative accuracy of the census and the A.C.E. Revision II estimates is based on an assessment of whether the confidence intervals for the evaluation poststrata cover 0 and $\hat{U}$.

The loss function analysis uses the estimated bias and variance to estimate an aggregate expected loss for the census and the A.C.E. Revision II for levels and shares for counties and places across the nation and within state. The loss function is the Mean Squared Error (MSE) weighted by the reciprocal of the census count for levels and the reciprocal of the census share for shares. The weight for both the census loss and the A.C.E. Revision II loss calculation is the reciprocal of the census count. The motivation for the selection of the groupings of areas for the loss functions is the potential use of the A.C.E. Revision II estimates in the postcensal estimates program.

Loss function analyses were carried for the following groups:

Levels

1. All Counties with population of 100,000 or less
2. All Counties with population greater than 100,000
3. All places with population at least 25,000 but less than 50,000
4. All places with population at least 50,000 but less than 100,000
5. All places with population greater than 100,000

State Shares

1. All Counties
2. All places

US Shares

1. All places with population at least 25,000 but less than 50,000
2. All places with population at least 50,000 but less than 100,000
3. All places with population greater than 100,000
4. All states

4. DATA REQUIREMENTS

1. We require the following files from DSSD:
   - Replicate coverage correction factors (CCFs) from DSE production variance estimation
   - Synthetic DSE estimates for states, counties, places using alternative duplication modeling assumptions
   - Direct DSE estimates for evaluation poststrata using alternative duplication modeling assumptions
   - Correlation bias estimates

2. We require the following files from PRED:
   - Replicate CCFs used in measurement bias estimation
   - Replicate measurement bias of CCFs
   - Replicate CCFs used for imputation variance estimation

3. We require the following files from DSCMO:
   - Poststratified micro-level census file

5. DIVISION RESPONSIBILITIES

1. The Decennial Statistical Studies Division (DSSD) will be responsible for managing the study.

2. The Planning, Research, and Evaluation Division (PRED), the Decennial Systems and Contract Management Office (DSCMO), and DSSD will be responsible for creating input files.

3. PRED, DSSD, and the Statistical Research Division (SRD) will collaborate to develop analysis plans, conduct the data analysis, and prepare the report.

4. See the milestone schedule for more specific information regarding responsibilities.

6. MILESTONE SCHEDULE

Activity	**Person(s) Responsible**	**Planned Start**	**Planned Finish**
Draft study plan	Mary	10/22/02	10/25/02
Computer design • Develop • Review	 Randy Mary, Bruce	 10/22/02 10/25/02	 10/30/02 11/3/02
Develop and test software: • Census tallies • Synthetic DSEs and sampling variances • Imputation variances • Bias estimates and variances • Duplication modeling variances • Loss functions • Confidence intervals	Randy	10/22/02	10/28/02

File deliveries: • Poststratified micro-level census file • Replicate CCFs for DSE production variances • Correlation bias estimates • Replicate CCFs for imputation variance estimation • Replicate CCFs for measurement bias estimation • Deliver alternative DSEs	 DSCMO Doug Eric Anne Katie Eric	12/15/02	12/15/02
Run software: • Census tallies • Synthetic DSEs and sampling variances • Imputation variances • Bias estimates and variances • Duplication modeling variances	Randy	12/16/02	12/23/02
Run loss function and confidence interval software	Randy	12/24/02	12/24/02
Deliver final results	Randy	12/24/02	12/24/02
Analysis Draft report Final report	Mary, Bruce, Randy	 12/25/02 12/31/02	 12/31/02 12/31/02

7. LIMITATIONS

- The estimated bias in the A.C.E. Revision II estimates may not account for all the sources of bias or may not account for the included nonsampling error components well. Estimates of correlation bias used in the A.C.E. Revision II are assumed to be without error.

- The estimated variance in the A.C.E. Revision II estimates may not account for all the sources of variance or may not account for the included nonsampling error components well, especially for error from choice of model for accounting for duplicates.

- The expected loss could instead have been measured by a loss function other than squared error weighted by the reciprocal of the census count.

9. RELATED STUDIES

- The Confidence Interval and Loss Function Analysis (CI&LF) will use data from the Census and Administrative Records Duplication Study to assess the error in the identification of census duplicates from the Further Study of Person Duplication (FSPD).

- CI&LF will use data from the Evaluation of the Error Due to Using Inmovers to Estimate Movers in the bias estimation.

- CI&LF will use data from the Evaluation of the Error Due to Inconsistent Poststratification Variables in the bias estimation

- CI&LF will use data from the Evaluation of the Missing Data Model in the variance estimation

- CI&LF will use data from the A.C.E. Revision II estimation programming to estimate the variance due to the choice of the duplication model in the variance estimation

9. REFERENCES

CAPE (1992) "Additional Research on Accuracy of Adjusted Versus Unadjusted 1990 Census Base for Use in Intercensal Estimates". Report of the Committee on Adjustment of Postcensal Estimates, Census Bureau, November 25, 1992.

Mulry, Mary (2002). "Chapter 7: Assessing the Estimates," in "A.C.E. Revision II: Design and Methdology." A.C.E. REVISION II MEMORANDUM SERIES #PP- 30. Census Bureau, Washington, DC.

Mulry, Mary H. and Spencer, Bruce D. (2001) "Overview of Total Error Modeling and Loss Function Analysis". DSSD Census 2000 Procedures and Operations Memorandum Series B-19* Census Bureau, Washington, DC.

Mulry, Mary H. and Spencer, Bruce D. (1993) "Accuracy of the of the 1990 Census and Undercount Adjustments". Journal of the American Statistical Association, 88, 1080-1091.

APPENDIX A

Estimating Bias in the Re ACE Estimates and Forming Confidence Intervals

Mary H. Mulry

This appendix describes a method for estimating the bias in the A.C.E. Revision II from four sources of error under the assumption that all other errors are zero, or at least negligible. The four sources of error are the error due inconsistency in the E-sample and P-sample reporting of the characteristics used in defining the poststrata, the error in identifying cases with census duplicates in both the E- and P-samples, the error due to using inmovers for outmovers in PES-C, and ratio estimator bias.

In addition, we describe the construction of confidence intervals for adjustment factors for estimation cells or aggregates of estimation cells, such as evaluation estimation cells.

We examine the errors with the current formulation of the match rate for the calculation of the A.C.E. Revision II presented in "Summary of A.C.E. Revision II Methodology" (Kostanich 2002).. We will use the same definitions of variables as found in the draft of Summary of A.C.E. Revision II Methodology.

First we discuss the correct enumeration rate and the match rate defined in Summary of A.C.E. Revision II Methodology. Then we discuss each of the four error components and develop how to estimate the bias from their combined effect.

Correct enumeration rate for A.C.E. Revision II

The correct enumeration rate for poststratum i for the calculation of the A.C.E. Revision II from Equation (5) of the draft of Chapter 6 is the following:

$$r_{CE,i} = \frac{CE_i^{ND} f_{i'}^{E} + \sum_{t \in i} W_{P,t}\, p_t\, z_t\, PR(CE)}{E_i}$$

where

E_i = total weighted E-sample in poststratum i.

CE_i^{ND} = correct enumerations without a census duplicate in poststratum i

$f_{i'}^{E}$ = double sampling adjustment for E-sample in Revision Sample poststratum i' . The Revision Sample poststrata are collapsed A.C.E. sample poststrata.

$PR(CE)$ = probability of that t is a correct enumeration (CEPROBF)

p_t = probability that enumeration t has a census duplicate outside the search area

z_t = probability that enumeration t with a census duplicate outside the search area is retained after unduplication (see draft of "Summary of A.C.E. Revision II Methodology")

$W_{P,t}$ = E-sample weight for person t.

For ease of discussion, we rewrite the correct enumeration rate for poststratum i as

$$r_{CE,i} = \frac{CE_i^{ND} f_{i'}^{E} + CE_i^{D}}{E_i}$$

where

$$CE_i^{D} = \sum_{t \in i} W_{P,t}\, p_t\, z_t\, PR(CE)$$

= correct enumerations with census duplicates in poststratum i .

Match rate for A.C.E. Revision II

The match rate for poststratum j for the calculation of the A.C.E. Revision II DSE from the draft of Summary of A.C.E. Revision II Methodology is the following:

$$r_{M,j} = \frac{M_{nm,j}^{ND} f_{j'}^{Mnm} + \frac{M_{om,j} f_{j'}^{Mom}}{P_{om,j} f_{j'}^{Pom}} [\, P_{im,j} f_{j'}^{Pim} + g_j \sum_{s \in j} W_{P,s}\, p_s\, (1-h_s)\, PR(res)\,] + \sum_{s \in j} W_{P,s}\, p_s\, h_s\, PR(res)\, PRm_{P,s}}{P_{nm,j}^{ND} f_{j'}^{Pnm} + [\, P_{im,j} f_{j'}^{Pim} + g_j \sum_{s \in j} W_{P,s}\, p_s\, (1-h_s)\, PR(res)\,] + \sum_{s \in j} W_{P,s}\, p_s\, h_s\, PR(res)}$$

where

$P_{nm,j}^{ND}$ = P-sample nonmovers without a census duplicate in poststratum j

$M_{nm,j}^{ND}$ = P-sample nonmover matches without a census duplicate in poststratum j

$P_{om,j}$ = P-sample outmovers in poststratum j

$M_{om,j}$ = P-sample outmover matches in poststratum j

$P_{im,j}$ = P-sample inmovers in poststratum j

$f_{j'}^{PG}$ = double sampling adjustment for P-sample group G, where G = nm, om, or im, in Revision Sample poststratum j' . The Revision Sample poststrata are collapsed A.C.E. sample poststrata.

$f_{j'}^{MG}$ = double sampling adjustment for matches in group G, where G = nm or om, in Revision Sample poststratum j'.

p_s = probability that person s has a census duplicate outside the search area
h_s = probability that person s with a census duplicate outside the search area is retained after unduplication (see "Summary of A.C.E. Revision II Methodology")

$W_{P,s}$ = P-sample weight for person s. The weight is assumed to include the probability of residence in draft Chapter 6, but that formulation needs to be reconsidered.

$PR(Res)$ = probability of being a resident of the sample block on Census Day (RPROB).

$PRm_{P,s}$ = probability person s with a census duplicate was matched in production
g_j = estimated proportion of P-sample persons in poststratum j with census duplicates outside the search area who are not retained as resident nonmovers by the duplicate study because they should have been coded as inmovers.

For ease of discussion, we rewrite the match rate for poststratum j as

$$r_{M,j} = \frac{M_{nm,j}^{ND} f_{j'}^{Mnm} + \frac{M_{om,j} f_{j'}^{Mom}}{P_{om,j} f_{j'}^{Pom}} \left[P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{D} \right] + M_{nm,j}^{D}}{P_{nm,j}^{ND} f_{j'}^{Pnm} + P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{D} + P_{nm,j}^{D}}$$

where

$P_{nm,j}^{D} = \sum_{s \in j} W_{P,s}\, p_s\, h_s\, Pr(Res)$ = P-sample nonmovers with census duplicates in poststratum j

$P_{nm-im,j}^{D} = g_j \sum_{s \in j} W_{P,s}\, p_s\, (1-h_s)\, Pr(Res)$ = P-sample nonmovers with census duplicates in poststratum j who are not retained as nonmovers by the duplicate study because they should have been coded as inmovers.

$M_{nm,j}^{D} = \sum_{s \in j} W_{P,s}\, p_s\, h_s\, PRm_{P,s}\, Pr(Res)$ = P-sample nonmover matches with census duplicates in poststratum j

Pr(Res = probability of being a resident (RBROB)

Corrections based on results of CARDS

Errors in the identification of census duplicates outside the search area may create a bias in the dual system estimate designed for the A.C.E. Revision II This bias affects the estimation of the E-sample correct enumeration rate and the P-sample match rate.

The Census and Administrative Records Duplication Study (CARDS) uses files created with administrative records to examine the effectiveness of the Further Study of Person Duplication in Census 2000 (FSPD) methodology. The FSPD refines the methodology for identifying and estimating the number of census duplicates. Using a computer matching algorithm, the study performs a match of the cases in the E-sample and the A.C.E. population sample, called the P-sample, to the census records for the entire nation. Links between the E-sample or the P-sample and the census enumerations are referred to as duplicates. CARDS first links the E and P samples to the administrative records and then attempts to confirm or deny duplicates identified by the FSPD. In addition, CARDS also identifies duplicates missed by the computer study, evaluates the FSPD process rules, and examines patterns of duplication.

The approach we are taking uses the results of the CARDS to correct the identification of cases with duplicates in the E- and P-sample for the A.C.E. Sample and the Revision Sample, a subsample of the A.C.E. Sample. We will use a research files for the E- and P-samples. For each case in the E- and P-sample with a FSPD census duplicate, CARDS will report it as correct, denied, or undetermined. CARDS also will identify cases in the E- and P-samples that have census duplicates not identified by the FSPD. With these results we will create new designations of cases with census duplicates, new values of the probabilities of having a census duplicate, and new values of the probabilities of cases with census duplicates being retained after unduplication.

First define new probabilities of having a census duplicate,

p_t^A = probability of census enumeration t has a census duplicate, corrected with data from administrative records,

p_s^A = probability of P-sample person s has a census duplicate, corrected with data from administrative records.

The steps for defining the corrected probabilities of having a census duplicate are as follows:

1. For census duplicates that CARDS denies:

- If in the E-sample, set $p_t^A = 0$, and if t matches s in the P-sample, set the corresponding $p_s^A = 0$.

- If in the P-sample, set $p_s^A = 0$, and if s matches t in the E-sample, set the corresponding $p_t^A = 0$.

2. For census duplicates that CARDS identifies, whether or not FSPD did:

- If in the E-sample, set $p_t^A = 1$, and if t matches s in the P-sample, set the corresponding $p_s^A = 1$.
- If in the P-sample, set $p_s^A = 1$, and if s matches t in the E-sample, set the corresponding $p_t^A = 1$.

3. For census duplicates CARDS can not determine and cases without duplicates:

- If in the E-sample, set $p_t^A = p_t$
- If in the P-sample, set $p_s^A = p_s$.

Next recalculate the new probabilities of P-sample person with a census duplicate being retained after unduplication, h_s^A, using the method described in Appendix 6.1 in draft Chapter 6 and the new set of cases with census duplicates and the new duplication probabilities. Also recalculate the new probabilities of E-sample enumeration with a census duplicate being retained after unduplication, z_t^A .

With the new p_t^A, p_s^A, z_t^A, and h_s^A, calculate a new correct enumeration rate and a new match rate. Let the superscript A denote a quantity calculated with the new p_t^A, p_s^A, z_t^A, and h_s^A .

$$r_{CE,i}^A = \frac{CE_i^{NDA} f_{i'}^{EA} + CE_i^{DA}}{E_i}$$

$$r_{Mj}^A = \frac{M_{nmj}^{NDA} f_{j'}^{MnmA} + \frac{M_{omj} f_{j'}^{Mom}}{P_{omj} f_{j'}^{Pom}} [P_{imj} f_{j'}^{Pim} + P_{nm-imj}^{DA}] + M_{nmj}^{DA}}{P_{nmj}^{NDA} f_{j'}^{PnmA} + P_{imj} f_{j'}^{Pim} + P_{nm-imj}^{DA} + P_{nmj}^{DA}} .$$

Correcting match rate for error due to using inmovers for outmovers

PES-C uses the number of inmovers to estimate the number of outmovers to avoid a bias caused by an underestimate of the number of movers. To examine the error caused by using the inmovers to represent outmovers, we rake the number of outmovers to total inmovers. The

distribution of the raked outmovers may better describe the outmovers than the distribution of the inmovers. Incorporating a correction in the match rate for using inmovers for outmovers requires defining:

$P_{im,j}^{O}$ = estimate of inmovers in poststratum j after raking the outmovers to the inmovers.

Correcting match rate for inconsistent poststratification variables

As discussed in "P-Sample match rate corrected for error due to inconsistent poststratification variables", inconsistency in the E-sample and P-sample reporting of the characteristics used in defining the poststrata may create a bias in the dual system estimate (DSE). This bias affects the estimation of the P-sample match rate.

The analysis for the A.C.E. Revision II will follow a similar investigation as for the original A.C.E. The basic approach is to estimate the inconsistency in the poststratification variables using the matches and then assume that the rates also held for the nonmatches. The models used for the inconsistency of the original A.C.E. poststrata ("Estimation of Inconsistent Poststratification in the 2000 A. C. E.", by Shelby J. Haberman and Bruce D. Spencer, 12/17/01) were fitted in two steps, first (i) models for inconsistency of basic variables, and then (ii) derivation of inconsistency probabilities for poststratification given the inconsistency probabilities of the basic variables. The inconsistency probabilities led to an estimate of the bias in the P-sample match rate that was used to estimate the bias in the DSE.

The approach we are taking for the Revised DSE is to calculate the proportions for the poststrata for the A.C.E. Sample. The proportions will not be applied in calculations of the double sampling adjustments based on the Revision Sample, a subsample of the A.C.E. Sample. We assume the models in (i) and (ii) have been revised to reflect revisions to the variables used in the P-sample poststratification. Incorporating a correction in the match rate for inconsistent poststratification variables requires defining:

$\hat{f}_G(j,k)$ = the proportion of group G persons enumerated in P-sample poststratum k who belong to P-sample poststratum j, based on their E-sample poststratification variables. The estimation of this proportion is based on the matched P-sample persons in group G. In this application, group G may be nonmovers, outmovers, or inmovers.

Next we need to define the following quantities:

$$P_{nm,j,I}^{NDA} = \sum_k \hat{f}_{nm}(j,k)\ P_{nm,k}^{NDA}$$

$$M_{nm,j,I}^{NDA} = \sum_k \hat{f}_{nm}(j,k)\ M_{nm,k}^{NDA}$$

$$P_{om,j,I} = \sum_k \hat{f}_{om}(j,k)\ P_{om,k}$$

$$P^{O}_{im,j,I} = \sum_k \hat{f}_{im}(j,k)\ P^{O}_{im,k}$$

$$M_{om,j,I} = \sum_k \hat{f}_{om}(j,k)\ M_{om,k}$$

$$P^{DA}_{G,j,I} = \sum_k \hat{f}_{nm}(j,k)\ P^{DA}_{G,j}\ , \text{ for G = nm or nm-im}$$

$$M^{NDA}_{nm,j,I} = \sum_k \hat{f}_{nm}(j,k)\ M^{NDA}_{nm,k}$$

Then we define the match rate corrected for the combination of error due to inconsistent poststratification variables, errors in identifying census duplicates, and error from using inmovers for outmovers, assuming no other errors are present, by the following:

$$r^{OA}_{M,j,I} = \frac{M^{NDA}_{nm,j,I} f^{MnmA}_{j'} + \dfrac{M_{om,j,I} f^{Mom}_{j'}}{P_{om,j,I} f^{Pom}_{j'}} \left[P^{O}_{im,j,I} f^{Pim}_{j'} + P^{DA}_{nm-im,j,I} \right] + M^{DA}_{nm,j,I}}{P^{NDA}_{nm,j,I} f^{PnmA}_{j'} + P^{O}_{im,j,I} f^{Pim}_{j'} + P^{DA}_{nm-im,j,I} + P^{DA}_{nm,j,I}}\ .$$

Calculation of Bias in the A.C.E. Revision II

From the draft of the "Summary of A.C.E. Revision II Methodology", the A.C.E. Revision II estimate for estimation cell ij formed by the intersection of E-sample poststratum i and P-sample poststratum j is

$$DSE_{ij} = Cen_{ij}\ (1 - r_{II,ij})\ \frac{r_{CE,i}}{r_{M,j}} CB_{ij}$$

where

$r_{II,ij} = (\ II_{ij} + LA_{ij}\ )\ /\ Cen_{ij}$, with II_{ij} as the census imputations, LA_{ij} as the late adds, Cen_{ij} as the census count including the late adds, and CB_{ij} the correlation-bias adjustment factor.

Then the bias in the A.C.E. Revision II estimate due to the combination of error due to inconsistent poststratification variables, errors in identifying census duplicates, and error from using inmovers for outmovers, assuming no other errors are present, for the estimation cell ij is given by

$$b_{ij,I}^{OA} = Cen_{ij}\ (1 - r_{II,ij})\ (\ \frac{r_{CE,i}}{r_{Mj}} - \frac{r_{CE,i}^{A}}{r_{Mj,I}^{OA}}\ )CB_{ij}\ .$$

When we add the correlation bias and the ratio estimator bias, we have the following bias estimate for the A.C.E. Revision II in estimation cell ij.

$$b_{ij} = b_{ij,I}^{OA} + b_{ij}^{CB} + b_{ij}^{R}\ .$$

[note: delete the middle term from the preceding equation]
where
b_{ij}^{R} = the ratio estimator bias in the A.C.E. Revision II in estimation cell ij .

The calculation of the ratio estimator bias makes use of the replicates formed for calculating the variance of the adjustment factors from the A.C.E. Revision II The calculation of b_{ij}^{R} is independent of the calculation of $b_{ij,I}^{OA}$ and hopefully will be a by-product of the variance calculations.

The bias in the adjustment factor for estimation cell ij is calculated by dividing the bias by the census count

$$b_{ij}^{*} = \frac{b_{ij}}{Cen_{ij}}$$

since the definition of the adjustment factor for estimation cell ij is

$$f_{ij}^{*} = \frac{DSE_{ij}}{Cen_{ij}}$$

Calculation of variance of bias in A.C.E. Revision II

The calculation of the variance of b_{ij} considers the variance of each of the three terms separately. For the loss function analysis we will assume that the variances of the correlation bias and the ratio estimator bias are zero. When constructing confidence intervals, we will assume that the multiplicative factor used to estimate correlation bias is a scalar and multiply the sampling variance by the square of the multiplicative factor. The calculation of the variance of $b_{ij,I}^{OA}$ will

use the 32 replicates of the E- and P-samples. These replicates were constructed by first partitioning the E- and P-samples into 32 groups and then removing a group. The replicate n is the whole sample with the nth group removed. For each replicate n, we will calculate $b_{ij,I,n}^{OA}$, n= 1,...,32, for each estimation cell ij. Then we will estimate the variance using a random group estimator as follows:

$$Var\ (\ b_{ij,I}^{OA}\ ) = \sum_{n}\ (\ b_{ij,I,n}^{OA} - b_{ij,I}^{OA}\ )^{2}\ .$$

We estimate the variance of the bias of the adjustment factor for estimation cell ij as follows:

$$Var\ (\ b_{ij,I}^{OA*}\ ) = \frac{Var\ (\ b_{ij,I}^{OA}\ )}{Cen_{ij}^{2}}$$

Forming confidence intervals

The confidence interval for the adjustment factor for the estimation cell ij for the A.C.E. Revision II estimate uses both the bias b_{ij}^{*} and the variance V_{ij}^{*} as follows:

$$(\ f_{ij}^{*} - b_{ij}^{*} - 2\sqrt{V_{ij}^{*}}\ ,\ f_{ij}^{*} - b_{ij}^{*} + 2\sqrt{V_{ij}^{*}}\ )$$

where

$$V_{ij}^{*} = (\ f_{ij}^{CB}\ )^{2}\ S_{ij}^{2} + V_{ij,M} + Var\ (\ b_{ij,I}^{OA*}\ )$$

S_{ij}^{2} = the sampling variance for the adjustment factor

f_{ij}^{CB} = the correlation bias correction factor

$V_{ij,M}$ = the variance due to missing data.

Confidence intervals for adjustment factors for aggregates of estimation cells, such as evaluation estimation cells, are defined using the same methodology. Estimates of the bias and variance as well as confidence intervals may be formed analogously for the undercount rate.

Reference

Kostanich, Donna (2002) "Summary of A.C.E. Revision II Methodology".DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-35. Census Bureau, Washington, DC.

APPENDIX B

Total Error Model and Loss Function Analysis

for the

A.C.E. Revision II Estimates of Population

Bruce D. Spencer

Draft 2.0

October 21, 2002

1. Overview

We consider estimation of expected loss when the loss functions are weighted sums of mean squared errors (MSEs) of the form $E(X_i - \theta_i)^2$, with i referring to the population of an area or other subgroup and X_i and θ_i referring either to population levels (numbers of people) or to shares (fraction of population in area or group i). The sums are taken over geographic areas, but the methodology extends without modification to arbitrary subgroups such as residents of geographic areas. Below, we will refer to areas.

The MSE is the sum of the variance and the squared bias. To estimate the squared bias, we need to allow for the variance in the estimate of bias, because the expectation of the square of a random quantity is equal to its variance plus the square of its mean. We show that, under certain conditions, the point estimate of difference in expected loss between the unadjusted

census and the adjusted census estimates does not need to incorporate an allowance for variances in the estimates of bias of the adjustment factors.

In this section, we provide an overview of logic of the analysis. In section 2 we describe the total error model and its implementation for the loss function analysis. Two different methods may be used to compute variances for use in the loss function analysis. One method, used in sections 2 and 3, below, is to store replicates; in some cases the replicates are based on sample replicates (as in jackknife or other pseudo-replication estimation of variance) and in other cases they are based on use of alternative models or assumptions. A second method, described in sections 4 and 5, is to use estimated variance-covariance matrices as the basis for deriving the variances and covariances needed in the loss function analysis. This method has been used in the past but it is somewhat cumbersome in the current situation, when the matrices may have dimensions of $10^4 \times 10^4$ or even greater. *I do not expect that the method described in sections 4 and 5 will be used for the loss function analysis of the A.C.E. Revision II estimates in 2002.* The description is included here for completeness.

To make clear the logic of the analysis, consider any single area or subgroup and let the following notation refer to population level or share, as the case may be. Here we suppress the subscript i for simplicity.

θ true quantity

C census count (unadjusted)

D adjusted estimate

υ $= \theta$ - C, net undercount

U = D - C is the estimate of υ

V_D variance of D

$\hat{V}_D$ estimate of V_D

β bias of D

B estimate of β

V_B variance of B

$\hat{V}_B$ estimate of V_B

V_{BD} covariance of B and D

$\hat{V}_{BD}$ estimate of V_{BD}

The MSE for the unadjusted census is υ^2, i.e., net undercount squared. If B is an unbiased estimator of β, then an unbiased estimate of υ is given by U - B. Recall that the expected value of the square of a random variable is equal to the sum of its variance and the square of its expectation. The variance of U - B is $V_D + V_B - 2V_{BD}$ and thus the expected value of $(U - B)^2$ is $\upsilon^2 + V_D + V_B - 2V_{BD}$. We therefore estimate the MSE for the unadjusted census by

(1) $$(U - B)^2 - (\hat{V}_B + \hat{V}_D - 2\hat{V}_{BD}).$$

The expected value of the estimator (1) is $[\upsilon - (EB - \beta)]^2 + (V_B - E(\hat{V}_B)) +$

$(E(\hat{V}_D) - V_D) - 2(E(\hat{V}_{BD}) - V_{BD})$. Thus, if B is an unbiased estimator of β and

$\hat{V}_B + \hat{V}_D - 2\hat{V}_{BD}$ is an unbiased estimator of $V_B + V_D - 2V_{BD}$, the estimator of MSE for the census given by (1) is unbiased. Note that if the covariance V_{BD} is estimated to be negligible, then $\hat{V}_{BD}$ drops out of (1), and the estimator of MSE of the unadjusted census simplifies to

(2) $$(U - B)^2 - (\hat{V}_B + \hat{V}_D).$$

The MSE for the adjusted estimate, D, is $\beta^2 + V_D$. To estimate this MSE we use

(3) $$B^2 - \hat{V}_B + \hat{V}_D,$$

which is unbiased if B is an unbiased estimator of β and $\hat{V}_D - \hat{V}_B$ is an unbiased estimator of $V_D - V_B$.

The excess MSE for the unadjusted census (C) relative to the adjusted estimate (D) is $\upsilon^2 - (\beta^2 + V_D)$, which we may estimate by (1) minus (3), or simply

(4) $$(U - B)^2 - B^2 - 2(\hat{V}_D - \hat{V}_{BD}).$$

If the covariance V_{BD} is estimated to be negligible, then we may use (2) instead of (1) and estimate the excess MSE by

(5) $$(U - B)^2 - B^2 - 2\hat{V}_D.$$

In this case, the point estimate for difference in expected loss between the adjusted and unadjusted census does not need to incorporate an allowance for $\hat{V}_B$. Previous loss function

analyses have assumed that V_{BD} would be relatively small and could be ignored. The method described in section 3, below, for estimating MSEs does not rely on an assumption that V_{BD} is negligible.

The following sections describe the calculation of expected loss in more detail. We use "area" as a general concept, and some care may be needed in practice. For example, if focusing numbers in a demographic group in an area, the area should be taken to exclude the groups not of interest.

2. Total Error Model

2.1. Overview

The following sources of error are considered in the total error model for the DSE:

a. Sampling variance (section 2.5.1)

b. Ratio-estimator bias (sections 2.5.2-2.5.3)

c. Bias due to inconsistency in the E-sample and P-sample reporting of the characteristics used in assigning the poststrata (sections 2.5.2-2.5.3)

d. Bias from error in identifying P-sample and E-sample cases that are duplicate enumerations in the census (sections 2.5.2-2.5.3)

e. Bias from error using inmovers for outmovers in PES-C (sections 2.5.2-2.5.3)

f. Correlation bias (sections 2.5.2-2.5.3)

g. Error due to choice of imputation model (section 2.5.4)

h. Error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" (section 2.5.5)

2.2. Input Variables from Production

The following variables are produced by the census and the A.C.E. for population estimation and are inputs for the loss function analysis. We note that the post-strata used for adjusting for erroneous enumerations and for undercoverage will not be the same; we use the term post-stratum to refer to an estimation cell that might involve a different E-sample and P-sample poststratum.

N_i census count (unadjusted estimate) for area i

H number of poststrata (or estimation cells)

H_E number of E-sample poststrata

H_P number of P-sample poststrata

C_{ih} census count for area i, poststratum h, against which adjustment factor is applied, $1 \le h \le H$

C^E_{ih} census count for area i, E-sample poststratum h, against which E-sample adjustment factor is applied, $1 \le h \le H_E$

C^P_{ih} census count for area i, P-sample poststratum h, against which P-sample adjustment factor is applied, $1 \le h \le H_P$

C_i vector of area i census counts across poststrata, to be multiplied by adjustment factors $= (C_{i1}, \ldots, C_{iH})^T$

C^E_i vector of area i census counts across E-sample poststrata, to be multiplied by E-sample adjustment factors

$= (C^E_{i1}, \ldots, C^E_{iHE})^T$

C^P_i vector of area i census counts across P-sample poststrata, to be multiplied by P-sample adjustment factors

$= (C^P_{i1}, \ldots, C^P_{iHP})^T$

f^E_j adjustment factor for E-sample poststratum j, $1 \le j \le H_E$. Note: these are assumed to include adjustments for II cases.

f^P_k adjustment factor for P-sample poststratum k, $1 \le k \le H_P$

f_h adjustment factor for poststratum h based on E-sample poststratum j and P-sample poststratum k

$= f^E_j f^P_k$

f^E vector of E-sample adjustment factors for E-sample poststrata

f^P vector of P-sample adjustment factors for P-sample poststrata

f vector of adjustment factors for poststrata

$= (f_1, \ldots, f_H)^T$

D_i adjusted estimate for area i

$= f^T C_i$

2.3. Bias Estimates

The following variables related to bias will be produced during the total error analysis.

b_h estimated bias in f_h

b estimated bias of f

$= (b_1, \ldots, b_H)^T$

B_i estimated bias in adjusted estimate for area i

$= b^T C_i$

T_i "target" estimate for area i

$= (f - b)^T C_i = D_i - B_i$

The vector b will be estimated as the sum of two components, $b = b_{meas} + b_{dup\text{-}modeling}$, with

b_{meas} estimate of net bias due to errors (b) - (f) in section 2.1.; see section 2.5.2,

$b_{dup\text{-}modeling}$ estimate of net bias from error (h) in section 2.1; see section 2.5.5.

2.4. Variance Estimates

Estimates of variance will be produced for D_i, T_i. and B_i. Two methods of estimation will be considered. Primarily, we will consider the use of replicates to develop the variance estimates for levels and shares. That method will avoid explicit use of a variance-covariance matrix. A second method will be described in section 4 that will explicitly use a variance-covariance matrix.

$\hat{V}_{sampling}(D_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.1

$\hat{V}(B_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.3

$\hat{V}_{sampling}(T_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.3

$\hat{V}_{imputation}(D_i)$ estimate of variance due to error (g) in section 2.1; see section 2.5.4

$\hat{V}_{dup-modeling}(D_i)$ estimate of variance due to error (h) in section 2.1; see section 2.5.5

The following overall variances are estimated as in section 2.5.6.

$$\hat{V}(D_i) = \hat{V}_{sampling}(D_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$$

$$\hat{V}(B_i) = \hat{V}_{sampling}(B_i)$$

$$\hat{V}(T_i) = \hat{V}_{sampling}(T_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i).$$

2.5. Summary Statistics Used to Develop Bias and Variance Estimates

2.5.1. Sampling Variance

The Census Bureau will prepare K replicates of the factors, based on sample replicates. It is possible that the factors may be vectors with full vector for f, or alternatively that the vectors will consist of a subvector of P-sample factors and a subvector of E-sample factors – the latter will involve less computer storage. The replicates of the factors may be used to compute variances of f and D_i's.

To generate the sampling variance of D_i, one would compute K estimates, say $D_{i(k)}$, $1 \leq k \leq K$, and derive the variance estimate accordingly, say $\hat{V}_{sampling}(D_i)$. (This same technique applies whether D refers to a population level or a share.)

2.5.2. Point Estimates of Bias Related to Data, Bias Related to Sampling, and Correlation Bias

There are a number of sources of bias in f. Mulry (2002) describes the estimation of error due to inconsistency in the E-sample and P-sample reporting of the characteristics used in assigning the poststrata, the error in identifying P-sample and E-sample cases that are duplicate enumerations in the census, error due to using inmovers for outmovers in PES-C, ratio estimator bias, and correlation bias. (If the A.C.E. Revision II estimates incorporate adjustments for correlation bias, the remaining correlation bias will be assumed to be negligible.) An estimate of b reflecting those sources of error will be produced by taking the difference between the production f and a version of f adjusted for the errors described earlier in this paragraph; the estimate will be denoted by b_{meas}.

2.5.3. Sampling Error of Point Estimates of Bias Related to Data and Sampling

A set of 32 sample replicates for use with the simple jackknife procedure has been developed at the Bureau by Katie Bench. Corresponding to each replicate, a value of f and a value of b_{meas} will be computed. The replicates may be used to compute sampling variances of b and T_i's. They may also be used to compute variances of f and D_i's, as an alternative to the replicates discussed in section 2.5.1 above. The replicates do not account for uncertainty in correlation bias estimates.

To generate the sampling variance of B_i and T_i, use the 32 replicates to develop $B_{i[k]}$, $D_{i[k]}$, and $T_{i[k]} = D_{i[k]} - B_{i[k]}$, $1 \leq k \leq 32$. Then derive the variance estimates accordingly, say

$\tilde{V}_{sampling}(B_i)$, $\tilde{V}_{sampling}(D_i)$, $\tilde{V}_{sampling}(T_i)$. We will ratio-adjust these in accordance with the variance estimate of D_i based on K replicates,

$$\hat{V}_{sampling}(T_i) = \lambda_i \tilde{V}_{sampling}(T_i) \text{ and } \hat{V}_{sampling}(B_i) = \lambda_i \tilde{V}_{sampling}(B_i),$$

with $\lambda_i = \hat{V}_{sampling}(D_i)/\tilde{V}_{sampling}(D_i)$. (The purpose of λ_i is to ratio-adjust the variance estimates for D_i using as controls the variance estimates based on larger numbers of replicates, while ensuring consistency among the sampling variance estimates for D_i, T_i, and B_i.)

2.5.4. Error from Choice of Imputation Model

Spencer (2002, section 3, step 7) provides for the construction of replicates that reflect error due to choice of imputation model. There are 128 replicates of vectors of factors, which we will denote by $f_{impute(k)}$, $1 \le k \le 128$. Note: it is assumed that the vectors of factors include adjustments for II cases.

To generate the variance of D_i from choice of imputation model, one would first compute 128 estimates of D_i, one from each of $f_{impute(k)}$, say $D_{i(k)}^{imp}$, $1 \le k \le 128$. For example, if $f_{impute(k)}$ is a vector referring to the adjustment factors and D_i refers to a population level for area i, one sets $D_{i(k)}^{imp} = (f^{impute(k)})^T C_i$ and if D_i refers to a population share, one sets

$D_{i(k)}^{imp} = (f^{impute(k)})^T C_i / \Sigma_j (f^{impute(k)})^T C_j$. One would then derive the variance estimate accordingly, say $\hat{V}_{imputation}(D_i) = \sum_{k=1}^{128} (D_{i(k)}^{imp} - \bar{D}_i^{imp})^2/127$, with $\bar{D}_i^{imp} = \sum_{k=1}^{128} D_{i(k)}^{imp}/128$. (This same technique applies whether D refers to a population level or a share.)

2.5.5. Failure of Assumptions in Modeling Duplication

An additional source of bias arises from error in the modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" in the A.C.E. Revision II estimates (Bell, Griffin, Kostanich, and Schindler 2002). To reflect this source of error, L alternative modeling assumptions will be used to generate alternative estimates of f, say $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$. (This is to occur during the production of the production factors, f.) The hypothetical bias due to modeling error in the production estimate when the alternative model k is true is

$b_{dup\text{-}modeling(\ell)} = f - f_{dup\text{-}modeling(\ell)}$.

It is reasonable to consider that no particular alternative model is correct, but still that the model used in the production estimate is incorrect. In this case, the modeling error may be treated as a random bias, in a manner similar to the treatment of the failure of the model underlying imputation of unresolved match, correct-enumeration, or residency status (Spencer 2002), and $b_{dup\text{-}modeling(\ell)}$ will be set to zero. To generate the variance of D_i from failure of assumptions for modeling duplication, one would take the L estimates, $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$, compute the corresponding L values of D_i, say $D_{i(\ell)}^{dup}$, $1 \le \ell \le L$, and derive the variance estimate

accordingly, say and derive the variance estimate accordingly, say

$\hat{V}_{dup-modeling}(D_i) = \sum_{\ell=1}^{L} (D_{i(\ell)}^{dup} - \bar{D}_i^{dup})^2/(L-1)$, with $\bar{D}_i^{dup} = \sum_{\ell=1}^{L} D_{i(\ell)}^{dup}/L$. (This same technique applies whether D refers to a population level or a share.)

2.5.6. Overall Variances

The overall variance of D_i is the sum of the sampling variance, variance from choice of models of duplication, and variance from choice of imputation model. Set

$\hat{V}(D_i) = \hat{V}_{sampling}(D_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$.

The variance of B_i is estimated as $\hat{V}(B_i) = \hat{V}_{sampling}(B_i)$.

The variance of T_i is estimated by the sum of the sampling variance of T_i, the variance in D_i from choice of models of duplication, and variance in D_i from choice of imputation model. Set $\hat{V}(T_i) = \hat{V}_{sampling}(T_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$.

3. Loss Function Calculations Based on Replicates

Aggregate loss functions for levels and shares are based on weighted sums and differences of $M_{C,i}$ and $M_{D,i}$. Define

$M_{C,i}$ Mean-Square Error (MSE) for Unadjusted Estimate of Level or Share for Area i

$M_{D,i}$ Mean-Square Error (MSE) for Adjusted Estimate of Level or Share for Area i.

$\hat{M}_{C,i} = (N_i - T_i)^2 - \hat{V}(T_i)$.

$$\hat{M}_{D,i} = B_i^2 + \hat{V}(D_i) - \hat{V}(B_i).$$

These estimates may be used to estimate the corresponding MSEs in loss functions.

4. Variance-Covariance Matrices

Define the following variance-covariance matrices for adjustment factors.

$\Sigma_{f,sampling}$ estimated sampling variance-covariance matrix of f, of dimension H × H

$\Sigma_{f,imputation}$ estimated variance-covariance matrix of f, of dimension H × H, reflecting error due to choice of imputation models for unresolved status.

$\Sigma_{f,dup\text{-}modeling}$ estimated variance-covariance matrix of f, of dimension H × H, reflecting error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities". It is possible that $\Sigma_{f,dup\text{-}modeling}$ will be set to zero.

Σ_f estimated variance-covariance matrix of f, of dimension H × H

$= \Sigma_{f,sampling} + \Sigma_{f,imputation} + \Sigma_{f,dup\text{-}modeling}$

It is possible that $\Sigma_{f,sampling}$ will be provided (e.g., by Douglas Olson), in which case Σ_f can be computed as described above. Alternatively, it may be estimated from the K replicates

described in section 2.5.1. As a final, although less precise method, $\Sigma_{f,sampling}$ may be estimable from the 32 replicates of f described in section 2.5.3.

The variance-covariance matrix for imputation error will be developed as described in Spencer (2002, section 2, step 7), with one possible modification. If the variance-covariance matrix described there is for the poststratum-level DSEs, the jk entry in the matrix must be divided by $C_{.j}C_{.k}$, with the subscript "." denoting national level census count for the poststratum. The matrix may also be estimated from the 128 replicates described in section 2.5.4.

The variance-covariance matrix $\Sigma_{f,dup\text{-}modeling}$ may be estimated from the vectors $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$, as described in section 2.5.5. Weighted moments may also be considered, if there is reason to consider weighting some of the alternatives more than others. We will set either (or both) of $\Sigma_{f,dup\text{-}modeling}$ and $b_{dup\text{-}modeling}$ to zero – model bias will be taken to be random (with mean zero) or fixed (with mean either non-zero or zero).

Sampling error, error due to choice of imputation model, and error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" are taken to be independent.

Define the following variance-covariance matrices involving bias estimates for adjustment factors.

Σ_b estimated variance-covariance matrix of b, of dimension H × H

Σ_{fb} estimated cross-covariance matrix of f and b, of dimension H × H

$$\Sigma = \begin{pmatrix} \Sigma_f & \Sigma_{fb} \\ \Sigma_{fb} & \Sigma_b \end{pmatrix}, \text{ of dimension } 2H \times 2H$$

Σ_{f-b} estimated variance-covariance matrix of f - b

$= \Sigma_f + \Sigma_b - 2\Sigma_{fb}$

The source for estimating Σ_b will be 32 sample replicates developed by Katie Bench. A collection of 32 values of b will be computed with respect to the replicates, and the variance-covariance matrix computed accordingly. At the same time, 32 values of f will be computed, one per replicate, and variance-covariance matrices $\Sigma_{f,sample}$ and Σ_{fb} will be computed. (If $\Sigma_{f,sample}$ was developed separately (e.g., by Douglas Olson), as described above, it may be desirable to compute a correlation matrix, say R_{fb} by from the replicate-based estimates, Σ_b, $\Sigma_{f,sample}$, and Σ_{fb}, and then to re-estimate Σ_{fb} by using the R_{fb} and Σ_b along with the original $\Sigma_{f,sample}$. The matrix R_{fb} will be of independent interest, because if its (off-diagonal entries) are small enough then we can use (2) and (5) instead of the more complicated (1) and (4).

5. Loss Function Analysis Based on Variance-Covariance Matrices

When we are estimating loss functions for shares based on variance-covariance matrices, the calculations are more complex than in section 3.

5.1. Loss Functions for Levels

Aggregate loss functions for levels are based on weighted sums and differences of $M_{C,i}$ and $M_{D,i}$. Now we use the following estimates for levels.

$$\hat{M}_{C,i} = (N_i - T_i)^2 - C_i^T \Sigma_{f-b} C_i$$

$$\hat{M}_{D,i} = B_i^2 + C_i^T \Sigma_f C_i - C_i^T \Sigma_b C_i$$

Aggregate loss functions for levels are based on weighted sums and differences of $M_{C,i}$ and $M_{A,i}$.

5.2. Loss Functions for Shares

Consider area i's share of aggregation G, where G is a union of areas. The unadjusted share is $N_{share,i} = N_i / \sum_{j \in G} N_j$. The adjusted share is $D_{share,i} = D_i / \sum_{j \in G} D_j$. The target share is $T_{share,i} = T_i / \sum_{j \in G} T_j$. The bias of the adjusted share is estimated by $B_{share,i} = D_{share,i} - T_{share,i}$.

5.2.1. Replications

For estimating variances of shares, we use replicates. As described below, only one set of Q replicates of f and b need to be generated. That set will service all of the loss functions for shares when the same specification of underlying variances is used. The value of Q is initially set at 1000.

Generate Q 2H×1 vectors $z^{(q)}$, $1 \le q \le Q$, from a multivariate normal distribution with mean zero and variance-covariance matrix Σ. Let the vector $x^{(q)}$ denote vector of the first H components of $z^{(q)}$ and let the vector $y^{(q)}$ denote the remaining H components of $z^{(q)}$; in other words, consider $z^{(q)}$ as a two H×1 vectors stacked on each other,

$$\mathbf{z}^{(q)} = \begin{pmatrix} \mathbf{x}^{(q)} \\ --- \\ \mathbf{y}^{(q)} \end{pmatrix}.$$

Observe that $x^{(q)}$ is distributed as the random error in f, $y^{(q)}$ is distributed as the random error in b, and the covariance between $x^{(q)}$ and $y^{(q)}$ is Σ_{fb}.

Define replicates of the adjustment factors f and bias estimates b by $f^{(q)} = f + x^{(q)}$ and $b^{(q)} = b + y^{(q)}$ for $1 \le q \le Q$. Replicates of adjusted estimates of shares and target values of shares are based on the replicates $f^{(q)}$ and $b^{(q)}$, and the variances and covariances are derived from the replicates.

Specifically, notice that the adjusted share for area i may be written as $D_{share,i} = \mathbf{f}^T\mathbf{C}_i / \sum_{j \in G} \mathbf{f}^T\mathbf{C}_j$. The q-th replicate of the adjusted share is

$$D^{(q)}_{share,i} = \mathbf{f}^{(q)T}\mathbf{C}_i / \sum_{j \in G} \mathbf{f}^{(q)T}\mathbf{C}_j.$$

The sample variance among the Q values of $D^{(r)}_{share,i}$ is used to estimate the variance of $D_{share,i}$. Denote the variance estimate by $V_{D,share,i}$.

Similarly, the q-th replicate of the target share is defined by

$$T^{(q)}_{share,i} = (\mathbf{f}^{(q)} - \mathbf{b}^{(q)})^T \mathbf{C}_i / \sum_{j \in G} (\mathbf{f}^{(q)} - \mathbf{b}^{(q)})^T \mathbf{C}_j .$$

The sample variance among the Q values of $T^{(q)}_{share,i}$ is used to estimate the variance of $T_{share,i}$.

Denote the variance estimate by $V_{T,share,i}$.

The q-th replicate of the bias in the adjusted share is defined by

$B^{(q)}_{share,i} = D^{(q)}_{share,i} - T^{(q)}_{share,i}$. The sample variance among the Q values of $B^{(q)}_{share,i}$ is used to estimate the variance of $B_{share,i}$. Denote the variance estimate by $V_{B,share,i}$.

5.2.3 MSEs for Shares

The MSE for the unadjusted estimate of share for area i is estimated by

$$\hat{M}_{C,share,i} = (N_{share,i} - T_{share,i})^2 - V_{T,share,i}$$

and the MSE for the adjusted estimate of share for area i is estimated by

$$\hat{M}_{D,share,i} = B_{share,i}^{\ 2} + V_{A,share,i} - V_{B,share,i}.$$

If there is zero correlation between f and b, then in $\hat{M}_{C,share,i}$ we may replace $V_{T,share,i}$ by the sum, $V_{D,share,i} + V_{B,share,i}$. The aggregate loss functions for shares are based on weighted sums and differences of $M_{C,share,i}$ and $M_{D,,sharei}$.

References

Bell, William R., Griffin, Richard A., Kostanich, Donna L., and Schindler, Eric (2002) Chapter 6: A.C.E. Revision II Estimation, Appendix 6.1. DRAFT of 09/19/02.

Mulry, M. H. (2002) Estimating Bias in the A.C.E. Revision II and Forming Confidence Intervals. 10/02/02 draft.

Spencer, Bruce D. (2002) Report on Missing Data Evaluation. October 5, 2002