ACCOUNTING FOR VARIANCE DUE TO IMPUTATION IN THE INTEGRATED COVERAGE MEASUREMENT SURVEY

Suzanne M. Dorinski, Richard Griffin[1]
Suzanne M. Dorinski, Bureau of the Census, Washington, DC 20233

Key words: jackknife; bootstrap

Introduction

Variance estimation methods used in post-enumeration surveys of previous censuses have not accounted for variance due to imputation. The unresolved cases in the P-sample have imputed probabilities of matching to the initial phase, while the unresolved cases in the E-sample have imputed probabilities of correct enumeration. In the Census 2000 Integrated Coverage Measurement (ICM) survey, the Census Bureau may impute probabilities for the enumeration or match status of unresolved cases and use a variance estimation method to account for the variance due to this imputation. We impute the probabilities by fitting hierarchical logistic regression models. This project compares three types of variance estimation: (1) a method developed by Schafer and Schenker (1991), (2) bootstrap, and (3) jackknife using the 1995 Census Test data for Oakland to determine which method is the best.

We use the 1995 production variance estimates (Fay and Town 1996) as the estimates of total sampling error. We then develop estimates of variance due to imputation for each of the three methods.

The next section briefly discusses the ICM process, the logistic regression programs used in imputation, and the resulting Dual System Estimates (DSEs). The succeeding sections describe the Schafer/Schenker, bootstrap and jackknife methods. Comparisons of the three methods follow. The final section presents conclusions.

Background

The Bureau of the Census will conduct Census 2000 with an unprecedented effort to count every resident in the United States. The effort will include:

- multiple mail contacts based on an improved mailing list for the initial phase,
- a toll-free telephone number at which callers can get answers to their questions as well as provide their Census response,
- blank forms at many convenient locations and in multiple languages, and
- a strong advertising and community-based publicity program.

In spite of the Bureau's best efforts, we will not be able to find every resident in the nation. The Bureau will conduct a second effort, known as the quality check or ICM, to determine what proportion of the population has been counted. The ICM in 2000 will be a nationally representative sample of 750,000 households.

The ICM is composed of the following steps:

independent listing of housing units,
housing unit matching,
ICM person interviewing,
DSE person matching, and
population estimation.

The independent listing of housing units is conducted in a sample of block clusters across the country. This independent listing is then matched to the mailing list used for the initial phase. The results of the housing unit matching are used to create an enhanced address list. This enhanced list is used to conduct ICM person interviewing in the sample block clusters.

The ICM person interviews are matched person-by-person to the results of the initial phase to determine the proportion of the population counted in the initial

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

phase. Unresolved cases are assigned an imputed probability of matching to the initial phase or an imputed probability of correct enumeration in the initial phase.

The 1995 ICM in Oakland consists of two samples: the P-sample and the E-sample. The E-sample consists of the original initial phase households in the ICM block clusters, while the P-sample consists of the independent rosters collected in the households in the ICM block clusters during the ICM interview. The 1995 Census Test also employs sampling for nonresponse follow-up in non-ICM blocks. The 1995 production variance estimates reflect error due to both types of sampling. For more information on the sample design, see Town and Fay (1995).

The logistic regression models include parameters for the following effects:

district office,
district office by ICM sampling stratum,
age,
tenure,
sex,
household size,
relationship to reference person,
amount of item imputation,
race,
sex by age,
race/ethnicity by age,
sex by race/ethnicity, and
race/ethnicity by sex by age.

The logistic regression model for the E-sample also includes parameters for the effects of structure, source of the data, and mail return. The logistic regression model for the P-sample also includes parameters for the effects of type of place, proxy interview, and outmover.

Cases with unresolved enumeration status can occur when the ICM interviewer is unable to obtain follow-up interviews with every household. In addition, interviewed households may have unresolved enumeration status when there is inadequate information available from the household interview. For example, the enumerators may only have been able to talk to a non-household member, or the person in question could have moved between the initial phase and the ICM interview but provided insufficient information to geocode the initial phase address. See Belin, et. al. (1992) and Diffendal and Belin (1991) for a more detailed discussion of unresolved cases in 1990. See table 1 for the 1995 imputation rates for each sample.

Belin, et. al. (1992) partition the variance due to imputation into three parts: (1) random errors in prediction given that the model is true and given that the estimated model parameters are the maximum likelihood estimates, (2) uncertainty due to the estimation of the model parameters given that the model is true, and (3) uncertainty in the model specification for the model parameters. This project focuses on the variance due to imputation from the first two sources. We assume that the model is specified correctly.

For all the methods, we are using DSEs. We produce variance estimates for the following 8 groups:

Black owners,
Black renters,
Asian Pacific Islander (API) owners,
API renters,
Hispanic owners,
Hispanic renters,
Other owners, and
Other renters.

The DSE formula in 1995 production is

$$DSE=(C_i-IIC_{NRFU,i})(\frac{CE_i}{E_i})(\frac{P_i}{M_i}),$$

where for poststratum i
C_i *is the initial phase count,*
$IIC_{NRFU,i}$ *is the number of whole person imputations,*
CE_i *is the weighted number of correct enumerations,*
E_i *is the weighted E-sample count,*
P_i *is the weighted P-sample count, and*
M_i *is the weighted number of matches.*

The imputed probabilities of correct enumeration are in the CE_i term, while the imputed probabilities of match are in the M_i term. If the case is resolved, a correct enumeration has a value of 1, while an incorrect enumeration has a value of 0. 1 indicates a match to the initial phase, while 0 indicates that the case does not match to the initial phase.

Schafer/Schenker method

The method developed by Schafer and Schenker (1991) to account for variance due to imputation is described in their paper. Their method is an analytic approximation to multiple imputation. To do the calculations, we use the imputed probabilities produced in 1995.

Schafer and Schenker regard the (C_i-$IIC_{NRFU,i}$) term as fixed, with the imputation variability occurring in the second and third terms. The imputation variance in the Schafer/Schenker method is C_1+C_2, where

$$C_1 = 2\left(\frac{\partial g(\hat{T})}{\partial T_Y}\right)^2 \sum_{i \in missing} w_i^2 \hat{\pi}_i (1-\hat{\pi}_i)$$

and

$$C_2 = \left(\frac{\partial g(\hat{T})}{\partial T_Y}\right)^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta}) \ ,$$

where

$$g(\hat{T}) = (C_i - IIC_{NRFU,i})\left(\frac{\sum_{j \in i} w_{E,j} Y_E}{\sum_{j \in i} w_{E,j}}\right)\left(\frac{\sum_{j \in i} w_{P,j}}{\sum_{j \in i} w_{P,j} Y_P}\right) \ ,$$

$$T_Y = \sum w_i Y_i,$$

$\hat{\pi}_i$ *is the imputed probability,*

$$D_\mu(\theta) = \sum_{i \in missing} w_i \left(\frac{\partial \mu_i(\theta)}{\partial \theta}\right),$$

parameters θ *are from the logistic regression models,*

$\mu(\hat{\theta})$ *are the imputed probabilities,*

and

Γ *is the variance estimate for the parameters of the logistic regression models.*

Separate logistic regression models are fit for each sample, so C_1 and C_2 are calculated separately for each sample. Since the models for the samples are independent, imputation variance is the sum of the C values from each sample.

As explained in Schafer and Schenker (1991), the C_2 term cannot be calculated directly because variance estimate Γ for the parameters of the logistic regression models are not readily available. We roughly estimate the component C_2 through bootstrap resampling. We use three E-sample and three P-sample bootstrap samples to measure the variation in the production DSEs given the ICM sample of blocks. Each bootstrap consists of selecting clusters with replacement. We refit the logistic regression models for each bootstrap sample to get different sets of model parameters. The different sets of model parameters are then used to impute for unresolved cases in the original production sample.

We produce 16 sets of DSEs, using the production P-sample and the results from the three P-sample bootstraps in all possible combinations with the production E-sample and the results from the three E-sample bootstraps. Since the sixteen samples are all possible pairings of four P- and four E-samples, there is a correlation structure among them that must be considered. We analyze the sixteen sets of DSEs as a 4x4 experiment with two random factors.

We express the DSEs under the different bootstrap samples as a linear model. Let P_i denote the P-sample used (0= production, 1-3 are the bootstrap alternatives), E_j denote the E-sample used (0 = production, 1-3 are the bootstrap alternatives), and DSE_{ij} denote the DSE for a poststratum group under P-sample i and E-sample j. Then DSEij = $\mu + P_i + E_j + PE_{ij}$, where μ is the grand mean, P_i is the main effect of P-sample i , E_j is the main effect of E-sample j, and PE_{ij} is the interaction of P-sample i with E-sample j. We treat the effects as random effects and estimate the components of variance due to each. Using analysis of variance, $C_2 = 1/3(MS_P + MS_E + MS_{PE})$, where MS denotes the mean squares due to each effect. For more details, see Mulry (1991) and Bateman (1991). See table 2 for the results of the Schafer/Schenker calculations.

Bootstrap

The bootstrap method to account for variance due to imputation is based on Shao and Sitter (1996). We resample the original data set 200 times to produce multiple data sets. To create each bootstrap sample, we resample with replacement the ICM block clusters in each ICM sampling strata, getting the same number of clusters in the resample as was originally observed in the sampling strata.

For each resampled data set, we impute probabilities for the enumeration status of unresolved cases (i.e. we refit the hierarchical logistic regression models for each resampled data set). Then we produce DSEs for each resampled data set, and use those estimates in the standard bootstrap formulas to estimate the variance due to imputation and ICM sampling.

For each resampled data set, we also produce DSEs using the 1995 production imputed values, and use those estimates in the standard bootstrap formulas to

estimate the variance due to ICM sampling. The difference between the two types of variance estimates is the variance due to imputation.

The variance estimates are given by

$$\frac{1}{200}\sum_{b=1}^{200}(\hat{\Theta}^{*b}-\bar{\Theta}^{*})^2$$

where

$\hat{\Theta}^{*b}$ *is the estimate of the bth bootstrap sample,*

$\bar{\Theta}^{*}=\frac{1}{200}\sum_{b=1}^{200}\hat{\Theta}^{*b}$.

Table 3 gives the results of the bootstrap calculations.

Jackknife

For the jackknife variance estimation, we refit the logistic regression models after deleting a cluster of the original data, then use the simple jackknife formula to estimate the variance due to imputation and ICM sampling. We also use the 1995 imputed probabilities and delete a cluster at a time to produce a jackknife estimate of the variance due to ICM sampling. The difference between the two types of variance estimates is the variance due to imputation.

We use the simple jackknife formula shown below

$$\frac{k-1}{k}\sum_{\alpha=1}^{k}(\hat{\Theta}_{(\alpha)}-\hat{\Theta}_{(.)})^2$$

where

k= number of clusters,

$\hat{\Theta}_{(\alpha)}$ *is the estimate excluding the kth cluster,*

$\hat{\Theta}_{(.)}=\sum_{\alpha=1}^{k}\frac{\hat{\Theta}_{(\alpha)}}{k}$

to calculate the variances. Since there are 161 clusters in the ICM sampling design, k=161.

Table 4 gives the results of the jackknife calculations. Note that the imputation variance estimate for Hispanic renters is negative. We would probably report the imputation variance in this case as 0. However, that might not be a reasonable estimate, since we impute about 15% of the P-sample data and about 16% of the E-sample data for Hispanic renters and would thus expect some variance due to imputation.

Results

Table 5 shows the imputation variance estimates produced by each of the methods. Overall results are mixed, with no one method clearly better than the others. For Black, Hispanic and API owners, the jackknife variance estimates are the smallest of the three methods. For Other owners, the bootstrap variance estimate is the smallest of the three methods. For the other estimates, the Schafer/Schenker estimate is the smallest of the three methods.

The logistic regression programs were run on a Sun Ultra 2 Model 2200 computer. The bootstrap processing was done on 2 data sets simultaneously. The 200 total data sets took 160.26 hours to complete. The jackknife processing was also done on 2 data sets simultaneously. The 161 total data sets took 155.72 hours to complete. The Schafer/Schenker method took about 10 hours to process, working on one data set at a time.

Recommendation

Nothing prevents the jackknife and bootstrap methods from producing negative estimates for imputation variance. The Schafer/Schenker method will produce positive estimates of imputation variance. Since we know that there should be some variance due to imputation, we recommend using the Schafer/Schenker method to calculate it.

References

Bateman, D. (1991) 1990 Coverage Studies and Evaluation Memorandum Series #A-9, "Final Report for 1990 PES Evaluation Project P1: Analysis of Reasonable Alternatives," dated July 9, 1991.

Belin, T.R., et. al. (1992) "Hierarchical Logistic-Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," Proceedings of the Eighth Annual Research Conference, U.S. Bureau of the Census, pp. 170-183.

Diffendal, G., and Belin, T. (1991) STSD Decennial Census Memorandum Series #V-112, "Results of

Procedures for Handling Noninterviews, Missing Characteristic Data, and Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey," dated July 1, 1991.

Fay, R.E. and Town, M.K. (1996) "Variance Estimation for the 1995 Census Test: Methodology and Findings," Proceedings of the Twelfth Annual Research Conference, U.S. Bureau of the Census, pp. 761-781.

Mulry, M. (1991) "1990 Post Enumeration Survey Evaluation Project P16: Total Error in PES Estimates for Evaluation Post Strata," dated July 11, 1991.

Schafer, J.L. and Schenker, N. (1991) "Variance Estimation with Imputed Means," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 696-701.

Shao, J. and Sitter, R.R. (1996) "Bootstrap for Imputed Survey Data," Journal of the American Statistical Association, Vol. 91, No. 435, pp. 1278-1288.

Town, M.K. and Fay, R.E. (1995) "Properties of Variance Estimators for the 1995 Census Test," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 724-729.

Acknowledgments

The authors wish to thank the Statistical Research Division of the Census Bureau for use of their computer for this project; Michael Ikeda for his suggestions during the planning phase of this project; and Leroy Bailey, Phil Gbur and Phil Steel for their comments on this paper.

Table 1. 1995 Match/Enumeration Status Imputation Rates for Oakland

Estimate	P-sample			E-sample		
	Imputed	Total	Rate	Imputed	Total	Rate
Black owners	250	3,150	7.94%	547	3,770	14.51%
Black renters	634	3,843	16.50%	1,040	5,028	20.68%
Hispanic owners	52	1,151	4.52%	113	1,112	10.16%
Hispanic renters	262	1,777	14.74%	340	2,073	16.40%
API owners	89	1,522	5.85%	151	1,703	8.87%
API renters	190	1,428	13.31%	227	1,781	12.75%
Other owners	255	3,941	6.47%	381	4,429	8.60%
Other renters	235	1,654	14.21%	349	2,121	16.45%

Table 2. Schafer/Schenker Results

Estimate	C_1	C_2	Imputation variance	1995 variance	Imputation as percent of 1995
Black owners	45,706.09	58,605.77	104,311.86	1,445,525.29	7.22%
Black renters	216,880.79	537,173.67	754,054.46	9,107,720.41	8.28%
Hispanic owners	16,461.93	7,134.84	23,596.77	496,179.36	4.76%
Hispanic renters	42,470.89	41,258.47	83,729.36	3,435,462.25	2.44%
API owners	14,180.25	6,796.07	20,976.32	819,025.00	2.56%
API renters	48,171.07	68,223.89	116,394.96	5,458,297.69	2.13%
Other owners	26,400.50	37,387.44	63,787.94	877,219.56	7.27%
Other renters	81,517.16	156,079.58	237,596.74	3,224,538.49	7.37%

Table 3. Bootstrap Results

Estimate	Bootstrap variance estimates			1995 variance	Imputation as percent of 1995
	With refitting	Without refitting	Imputation variance		
Black owners	1,911,352.33	1,761,109.21	150,243.12	1,445,525.29	10.39%
Black renters	14,126,954.96	12,313,268.30	1,813,686.66	9,107,720.41	19.91%
Hispanic owners	734,146.73	669,279.92	64,866.81	496,179.36	13.07%
Hispanic renters	5,706,180.30	5,605,055.48	101,124.82	3,435,462.25	2.94%
API owners	572,234.24	521,266.00	50,968.24	819,025.00	6.22%
API renters	5,044,268.83	4,766,813.06	277,455.77	5,458,297.69	5.08%
Other owners	674,097.84	655,557.92	18,539.92	877,219.56	2.11%
Other renters	2,620,668.07	2,220,326.09	400,341.98	3,224,538.49	12.42%

Table 4. Jackknife Results

Estimate	Jackknife variance estimates			1995 variance	Imputation as percent of 1995
	With refitting	Without refitting	Imputation variance		
Black owners	1,356,966.14	1,274,661.29	82,304.85	1,445,525.29	5.69%
Black renters	13,237,203.79	10,704,406.29	2,532,797.50	9,107,720.41	27.81%
Hispanic owners	452,078.39	449,409.19	2,669.20	496,179.36	0.54%
Hispanic renters	3,890,460.72	4,110,752.96	-220,292.24	3,435,462.25	N.A.
API owners	706,834.17	693,865.88	12,968.29	819,025.00	1.58%
API renters	4,863,005.69	4,513,405.81	349,599.88	5,458,297.69	6.40%
Other owners	898,625.91	829,091.97	69,533.94	877,219.56	7.93%
Other renters	3,578,621.53	3,229,214.38	349,407.15	3,224,538.49	10.84%

N.A. -- not applicable

Table 5. Comparison of Methods

Estimate	Imputation variance estimates			Imputation as percent of 1995 variance		
	Schafer/ Schenker	Bootstrap	Jackknife	Schafer/ Schenker	Bootstrap	Jackknife
Black owners	104,311.86	150,243.12	82,304.85	7.22%	10.39%	5.69%
Black renters	754,054.46	1,813,686.66	2,532,797.50	8.28%	19.91%	27.81%
Hispanic owners	23,596.77	64,866.81	2,669.20	4.76%	13.07%	0.54%
Hispanic renters	83,729.36	101,124.82	-220,292.24	2.44%	2.94%	N.A.
API owners	20,976.32	50,968.24	12,968.29	2.56%	6.22%	1.58%
API renters	116,394.96	277,455.77	349,599.88	2.13%	5.08%	6.40%
Other owners	63,787.94	18,539.92	69,533.94	7.27%	2.11%	7.93%
Other renters	237,596.74	400,341.98	349,407.15	7.37%	12.42%	10.84%

N.A. -- not applicable