A STUDY IN HETEROGENEITY OF CENSUS COVERAGE ERROR FOR SMALL AREAS

Mary H. Mulry, The M/A/R/C Group, and Mary C. Davis, and Joan M. Hill*, Bureau of the Census
Mary H. Mulry, The M/A/R/C Group, 7850 North Beltline Road, Irving,TX 75063

Keywords: logistic regression; undercount; dual system estimation

1. Introduction

Estimates of coverage error for the 1990 Census were based on dual system estimation (DSE) where one system was the census enumeration, and the second enumeration was done for a sample of the population, the P Sample, as part of the Post Enumeration Survey(PES). The PES also had a sample of census enumerations, the E Sample, which estimated erroneous inclusions in the census. Using these two samples and the census, estimates of population size were made DSE and subsequently used to estimate the coverage of the census.

An assumption underlying DSE for census coverage error is that the capture probabilities for the Census or the P Sample are equal. Since it is obvious that the capture probabilities are not uniform for all members of the population, the Census Bureau application forms poststrata based on variables which previous studies have shown to be correlated with coverage error such as tenure, race and Hispanic ethnicity, age, sex, and urbanization. Then the estimation assumes that the capture probabilities are uniform within these poststrata. Certainly the poststrata improve the estimation over what would be achieved without it. However several studies using the 1990 PES data found evidence to suggest that heterogeneity in capture probabilities remained within the poststrata. (Hentgartner and Speed 1993 and Alho, Mulry, Wurdeman, and Kim 1993)

To obtain estimates for areas within the poststrata, the DSE is distributed to blocks proportional to the size of the poststratum's population within the block. This method is known as synthetic estimation. It too assumes that the capture probabilities are uniform within the poststrata. Another way of saying this is that the coverage error rate is equal for the portion of the poststratum's population in each block.

Remaining heterogeneity in capture probabilities within a poststratum effects the estimates in two ways: (1) it makes the poststratum estimate too low (also called correlation bias), and (2) synthetic estimation within poststratum does not capture the variation in coverage error for small areas.

The Census Bureau's Committee on Adjustment of Postcensal Estimates (1992) investigated the quality of small area estimation with a revision of the 1990 PES estimates. They concluded that, on average, the adjusted state numbers were more accurate than the unadjusted state numbers. However, no conclusions were reached as to whether the adjusted numbers for smaller areas were an improvement over the census numbers.

In other work, Thompson (1994) shows that the synthetic method for small area estimation does not correct for variation in census coverage error at the block level but neither is it worse than the census. Another reference for similar findings is Schindler and Navarro (1995).

The goal of the study is to investigate whether it is possible to improve upon synthetic estimation for small areas. Although correlation bias is a very important and related topic, it will not be covered. The focus is on analyzing the heterogeneity in census coverage error at the block level and evaluating alternative ways of estimating it.

The approach uses estimates from a generalized dual system estimator based on logistic regression. Basically the method is a form of synthetic estimation where the DSE is distributed within a poststratum proportional to the generalized DSE for a block instead of its census count.

More specifically, this paper investigates the feasibility of using estimates of the probability of a person being enumerated in the census in developing models of the heterogeneity in census coverage error for small areas. Revisions of logistic regression models for these probabilities (Alho, Mulry, Wurdeman, and Kim 1993) are developed using data from the 1990 census and post enumeration survey (PES). The independent variables in these models are characteristics of the person, their household, and their block derived from the short form data without using any of the characteristics of the census or PES. The probabilities may be used to develop estimates of coverage error for small areas.

The paper contains a description of the methodology for block level estimation followed by its evaluation. The results are contained in Section 4 with the final section containing a summary.

2. Methodology

The analysis will evaluate the quality of several alternative methods for estimating census coverage error at the block level by assessing how well these methods estimate the heterogeneity in census coverage error at the block level. The analysis considers only the 1990 PES block clusters and unweighted data from the Post Census Review (PCR) estimates published in 1992.

The way of estimating the number of people in a block under consideration is calculated by using the estimated probabilities of being included in the census p_{1bi} and the estimated probabilities of being included in the P-sample population p_{2bi} based on conditional logistic regression models developed using data from the 1990 Post Enumeration Survey (Alho, Mulry, Wurdeman, and Kim, 1993). The explanatory variables are based on data from the 1990 Census short form related to the individual, the individual's block, and geography.

Four separate models are fit for minorities and nonminorities, owners and renters, in urban areas with population of 250,000 or more. These groups correspond to evaluation poststrata used in the evaluation of the 1990 PES and are aggregates of PES poststrata. The strategy is to first examine these areas and if the results are positive, continue to investigate the method for the rest of the poststrata. For each block, we will consider several estimators, each distribute the estimate $\hat{N}_{DSE_k}$, for poststratum k, k = 1,...K, in different ways. Notice that the estimator for $\hat{N}_{jb}$ is of the same form as the synthetic estimator used for the 1990 PES with $\hat{N}_{Sjk}$ replaced by the census count for the poststratum and $\hat{N}_{jkb}$ replaced by the poststratum's census count in the block. The basic formula for the estimate in block b, where b = 1,...,B, is

$$\hat{N}_{jb} = \sum_{k=1}^{K} f_{jk} \hat{N}_{jkb}$$

where j = 1, 2 and

$$f_{jk} = \hat{N}_{DSE_k} / \hat{N}_{Sjk}$$

$$\hat{N}_{jkb} = \sum_{M_{bi} r_{bi}=1} 1/\phi_{bi} + \sum_{n_{jbi}(1-r_{bi})=1} \gamma_{jbi} / p_{jbi}$$

where

j = 1 if the estimate is calculated using resolved cases and the unresolved E-sample cases,
2 if the estimate is calculated using resolved cases and the unresolved P-sample cases

p_{1bi} = estimated probability of the i-th person in the b-th block being included in the census enumeration

p_{2bi} = estimated probabilities of the i-th person in the b-th block being included in the P-sample population

$\phi_{bi} = p_{1bi} + p_{2bi} - p_{1bi} p_{2bi}$
= the probability of the i-th person in the b-th block being included at least once.

γ_{jbi} = the imputed probability of the i-th person in t he b-th block being enumerated from the PES imputation,

r_{bi} = 1 if the i-th individual's case in block b is resolved
0 otherwise

M_{bi} = 1 if the i-th individual in block b is included in the P sample or E sample or both,
0 otherwise

n_{jbi} =1 if the i-th individual in block b is included in the j-th sample,
0 otherwise

M_{bi} and n_{jbi} are determined in the following manner:

$$M_{bi} = u_{1bi} + u_{2bi} + m_{bi}, \quad n_{jbi} = u_{jbi} + m_{bi}$$

where

u_{1bi} = 1 if the i-th individual in block b is included in the census enumeration and not in the P-sample population,
0 otherwise.

u_{2bi} = 1 if the i-th individual in block b is included in the P-sample population and not in the census enumeration,
0 otherwise.

m_{bi} = 1 if the i-th individual in block b is included in both the census enumeration and the P-sample population,
0 otherwise.

The estimate of the population $\hat{N}_{Sjk}$ may be calculated using data from the Post Enumeration

Survey with the following estimator, suppressing the k subscript for the k-th poststratum:

$$\hat{N}_{Sj} = \sum_{b=1}^{B} \sum_{M_{bi} r_{bi}=1} W_{bi}(1/\phi_{bi})$$

$$+ \sum_{b=1}^{B} \sum_{n_{jbi}(1-r_{bi})=1} W_{bi}(\gamma_{jbi}/p_{jbi})$$

for j = 1,2 where

W_{bi} = survey weight for the i-th person in block b.

For use in the estimation which follows, the p_{2bi} estimated by the logistic regression model are multiplied by the nonmover rate in an adjustment cell. For the remainder of this paper, all P-sample inclusion probabilities are post-adjustment p_{2bi}'s.

3. Evaluation

The evaluation of the alternative estimates for blocks is difficult because the true value of the population size for the blocks can not be known. The approach that this study takes is to consider six measures of the population of a block considered standards of comparison. None of the standards is perfect, but each has something to offer. The assessment will have to weigh all of them in order to reach any conclusions. Three standards defined below, the S_{jb}, for j= 0, 1, 2, are based on the people seen in both the P Sample and the census combined with variations in assumptions about the unresolved cases. The advantage is that they are based solely on data without any modeling, except for the imputation models. The disadvantage is that they do not include people seen in neither list. The other three standards, $\tilde{N}_{jb}$, for j= 0, 1, 2, are based on the model underlying dual system estimation and the logistic regression model of inclusion probabilities which assume that the models are appropriate and fit the data well. However, these standards do account for people not captured in either list.

The first three measures of the population in block b use unweighted data to determine the correct number of people in a block according to PES by calculating the sum S_{jb}, for j= 0, 1, 2, of the nonmover matches, the other correct enumerations, the nonmover nonmatches, and the sum of probabilities of being enumerated for the nonmover unresolved cases and the cases that have an unresolved status because it could not be determined whether they were movers.

$$S_{0b} = M_b + CE_b + NM_b \sum_{n_{1bi}(1-r_{bi})=1} \gamma_{1bi} + \sum_{n_{2bi}(1-r_{bi})=1} \gamma_{2bi}$$

$$S_{1b} = M_b + CE_b + NM_b + \sum_{n_{1bi}(1-r_{bi})=1} \gamma_{1bi}$$

$$S_{2b} = M_b + CE_b + NM_b + \sum_{n_{2bi}(1-r_{bi})=1} \gamma_{2bi}$$

The sums S_{jb} for j= 1, 2, are biased downward because they do not include the people not included in either list. However, sum S_{0b} is based on the unresolved cases from both the P and E samples. There is some chance that the unresolved cases are the same people so the other two estimators are based on only the unresolved cases in the P sample or the E sample. This factor makes it unclear whether S_{0b} is biased and the direction of the bias. The evaluation explores the influence of the different combinations of unresolved cases.

There are two concerns about using the sums S_{jb} as standards. One is that they may not account for noninterviews in the P sample. To the extent that housing units that were not interviewed in the P sample were enumerated in the census, noninterviews in P sample are not a problem. However, there are some housing units where data that the PES process could use was not collected in either system. If the data files used in the study had housing unit coverage data on them, possibly some of the misses could be unraveled. The results of this study should be valid as a feasibility study without this data. However, linking the housing unit coverage information and the person coverage information is beyond the scope of this project, but may be conducted if deemed desirable after reviewing these results.

The other concern is that the sums S_{jb} do not include any outmovers who were missed by the census. The definition of the P-sample population included inmovers and not outmovers. The E-sample cases which are designated as correct enumerations do not have information as to whether they were outmovers. Therefore, an adjustment for the missed outmovers is difficult to calculate, particularly without housing unit coverage data.

The other three standards are calculated by using the estimated probabilities of being included in the census p_{1bi} and the estimated probabilities of being included in the P-sample population p_{2bi} based on conditional logistic regression models developed using data from the 1990 Post Enumeration Survey (Alho, Mulry, Wurdeman, and Kim, 1993). The estimator uses the fact that the probability of being included at least once for the i-th person in block b can be estimated by

$$\phi_{bi} = p_{1bi} + p_{2bi} - p_{1bi} p_{2bi}$$

With the models appropriate for an area and the census enumerations for the area, the probabilities of inclusion can be estimated for each enumeration. In this case, imputed enumerations are included along with enumerations based on data obtained from interviews. Since the imputations are created by a hot-deck procedure, the models are assumed to work as well on imputations as they do on the other enumerations.

If all the enumerations for the census in a block are correct, an estimate of the population can be obtained by the following Horvitz-Thompson estimator,

$$\tilde{N} = \sum_{g} 1/\phi_g$$

However, not all census enumerations are correct, as has been documented by the E Sample of the PES. The alternative is to modify this estimator to account for the possibility of erroneous enumerations. The approach we take is first to estimate the probability of an enumeration being correct with the correct enumeration rate in the bi-th individual's PES poststrata, c_k. Again, the c_k are assumed to hold for the hot-deck imputations although they were calculated using only enumerations with data from interviews. Then an estimator, which we will use as a standard, is as follows:

$$\tilde{N}_{0b} = \sum_{k=1}^{K} \sum_{i=1}^{nkb} c_k (1/\phi_{kbi})$$

Two alternative estimators of the population size, which are also used as standards are defined below. One is based only on the probability of being included in the P-sample population, and the other uses only the probability of being included in the census.

$$\tilde{N}_{1b} = \sum_{k=1}^{K} \sum_{i=1}^{nkb} c_k (1/p_{1kbi})$$

$$\tilde{N}_{2b} = \sum_{k=1}^{K} \sum_{i=1}^{nkb} c_k (1/p_{2kbi})$$

Now that we have developed the six standards, we describe how they are used in the evaluation. The evaluation focuses on the difference at the block level between the estimators $\hat{N}_{jb}$ for j=1, 2, with standards defined by evaluation estimators S_{jb}, for j= 0, 1, 2, and $\tilde{N}_{jb}$ for j= 0, 1, 2. The analysis consists of examining the level of error and the relative error. For example, the difference between $\hat{N}_{1b}$ and a standard is defined by

$$D_b = \hat{N}_{1b} - T_b$$

where T_b is one of the standards $\tilde{N}_{jb}$ or S_{jb} ,for a total of six differences. Then the relative difference is defined by

$$RD_b = (\hat{N}_{1b} - T_b)/T_b$$

The same calculations are made for $\hat{N}_{2b}$.

Also, the census count, $\hat{N}_{CEN,b}$, and the synthetic estimates from the PES, $\hat{N}_{syn,b}$, are compared with the standards. The distribution of the errors, the average error and average relative error, and the range of the error and range of relative error are calculated

4. Results

The evaluation is restricted to blocks in the PES sample that are contained in the area covered by the four models which are minorities and nonminorities, owners and renters, in urban areas with population of 250,000 or more.. Table 4 contains the covariates, coefficients, and standard errors for the logistic regression models of census inclusion probabilities models for minorities. Space does not permit displaying the other models.

The results of the comparisons of estimators with the standard S_{0b} are contained in Tables 1 through 3. The results for the three different estimators must be viewed relative to each other and relative to the standard. These results do not represent calculations of what the true errors in the census count for these blocks are. The comparisons with the other two standards show similar patterns. Also, the results for $\hat{N}_{jb}$ for j=1, 2 are similar.

Since estimators may perform differently in blocks where there are few errors than in blocks where there are many errors, tables are made for blocks grouped in error categories of small and large error, both positive and negative. The cut-off points for the categories are determined by the difference in the E-sample total and the S_{0b} for the blocks. The categories used are E-sample count being more than 3 people too low, the difference being between 3 people too many and 3 people too few, and the E Sample being more that 3 people too high. These seemed to be natural breaks when viewing the distribution.

So that a few extreme errors do not distort the calculations of average error and average relative error, the distributions are trimmed by discarding the five largest negative errors and the five largest positive errors.

When all the blocks are pooled together, both estimators $\hat{N}_{jb}$ for j=1, 2, exhibit reduced average error, median error, and range of error when compared with the three standards S_{jb} for j= 0, 1, 2. The results with S_{jb} for j= 1, 2 follow a similar pattern as the results for S_{0b} which we will discuss. With the trimmed distributions, the average error relative to the standard S_{0b} for $\hat{N}_{jb}$ for j=1, 2 is -0.06 and -0.04 respectively. The median error for the two estimators is -0.46 and -0.47 which is an error of approximately one-half person too low. The average errors for the census count, $\hat{N}_{CEN,b}$, and the synthetic estimates from the PES, $\hat{N}_{syn,b}$, are 0.25 and 1.70, respectively. The median errors are 1.00 and 1.44, respectively.

As for relative error when the standard is S_{0b}, the average relative errors for $\hat{N}_{jb}$,for j=1, 2, are almost negligible at 0.003 and 0.005, respectively. The median relative errors are -0.014 and -0.015 respectively. The average relative errors for the census count, $\hat{N}_{CEN,b}$, and the synthetic estimates from the PES, $\hat{N}_{syn,b}$, both equal 0.045. The median errors are 0.025 and 0.031, respectively. The range of relative error for $\hat{N}_{jb}$, for j=1, 2 are reduced over what they are for $\hat{N}_{CEN,b}$ and $\hat{N}_{syn,b}$

When viewing the results for the blocks with small errors, there is not much difference in the estimators. However, the estimators $\hat{N}_{jb}$ for j=1, 2 appear much more effective for the blocks with the larger coverage errors, both positive and negative. Interestingly, when there is a large positive error ("overcount"), the synthetic estimator appears to increase the amount of error.

5. Summary

The results for the first three standards appear positive but no definitive judgments can be made until comparisons with the other three standards can be made. This work is in progress. We emphasize that the results for the estimators must be viewed in a context relative to each other and not in absolute terms.

The $\hat{N}_{jb}$ for j=1, 2 are not the form an estimator which can be used in the census since erroneous enumerations only will be determined for the blocks in the PES. A version which can be estimated would use the form of the estimator used for $\tilde{N}_{jb}$ for j= 0, 1, 2. The denominators for the adjustment factors based on the sample in this paper would be replaced by the estimators using the whole census file for the urban areas with population of 250,000 or more. We did not consider these estimators in this study mainly because of the time required in obtaining the whole census file. With this version of the estimator, the DSE for the poststrata would be the same as the one used in the 1990 PES. However, the distribution of the DSE within the poststrata would be proportional to the generalized DSE, not the census as is done with synthetic estimation..

Acknowledgment

The authors want to thank Henry Woltman and Bruce Spencer for reviewing the manuscript and offering helpful suggestions.

*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.

References

Alho, J. M., Mulry, M. H., Wurdeman, K., Kim, J. (1993) "Estimating Heterogeneity in the Probabilities of Enumeration for Dual System Estimation," Journal of the American Statistical Association, 88, 1130-1136.

Committee on the Adjustment of Postcensal Estimates (1992) "Assessment of the Accuracy of Adjusted versus Unadjusted 1990 Census Base for Use in Intercensal Estimates." Unpublished manuscript, August 7, 1992. U. S. Bureau of the Census.

Hentgardner and Speed, T. (1993) "Assessing Between-Block Heterogeneity Within the Post-strata of the 1990 Post-Enumeration Survey." Journal of the American Statistical Association, 88,1119-1125.

Schindler, E. and Navarro, A. (1995) "Effect of Sampling for Nonresponse Followup in the Census Environment on Population Estimates," Proceedings of the 1995 Annual Research Conference, U. S. Bureau of the Census.

Thompson, John (1994) "Effects of Sampling for Nonresponse Follow-up and Integrated Coverage Measurement on Block Cluster Level Estimates." DSSD 1995 Census Test Memorandum Series, #D-1, Decennial Statistical Studies Division, Bureau of the Census.

Table 1. Small "Coverage Error", 1040 blocks

	Ave. Error	Range of Error	Ave. Rel. Error	Range of Relative Error
$\hat{N}_{1b}$	0.76	-40,37	-0.005	-0.75,0.90
Synthetic	-1.34	-5,20	0.033	-0.51,1.06
Census	-0.49	-2,15	0.016	-0.50,1.00

Table 2. Large "Undercount", 550 blocks

	Ave. Error	Range of Error	Ave. Rel. Error	Range of Relative Error
$\hat{N}_{1b}$	0.49	-63,51	0.005	-0.42,0.80
Synthetic	-8.11	-76,44	-0.126	-0.93,0.35
Census	-10.26	-77,35	-0.148	-0.93,0.28

Table 3. Large "Overcount", 680 blocks

	Ave. Error	Range of Error	Ave. Rel. Error	Range of Relative Error
$\hat{N}_{1b}$	0.57	-56, 55	0.014	-0.830, 0.84
Synthetic	10.18	0.6, 69	0.200	0.003, 4.41
Census	8.34	3, 63	0.175	0.016, 4.40

Table 4. Logistic Regression Models for Census Inclusion Probabilities for Minorities in Urban Areas with Population 250,000 or more

	Owner	Renter
Intercept	2.552 (0.101)	1.477 (0.081)
Age	0.079 (0.057)	-0.084(0.038)
Sex	0.164(0.036)	0.212 (0.028)
Black/Non-Black	-0.238 (0.089)	0.077 (0.074)
Hispanic	0.172 (0.083)	0.170 (0.70)
Marital Status	0.356 (0.051)	0.099 (0.041)
Household size	-0.218 (0.018)	0.014 (0.016)
% non-owner	-0.126 (0.025)	0.148 (0.021)
% Black or non-Hispanic	-0.095 (0.027)	-0.214 (0.019)
% multi-units	-0.045 (0.030)	-0.160 (0.022)
Vacancy Rate	-0.065(0.017)	-0.111(0.013)
Age*Black/non-Black	0.177(0.043)	0.087 (0.031)
%non-owner* %Black/non-Black		-0.031 (0.021)
Metropolitan place	-0.035 (0.040)	-0.079 (0.031)
Metropolitan* Black/non-Black	0.124 (0.023)	0.065 (0.021)
age^2	0.124(0.023)	0.065 (0.021)
age^3	-0.042 (0.017)	0.016(0.011)
age*sex	0.093 (0.038)	0.137(0.029)
not related to person 1	-0.966 (0.077)	-0.773 (0.053)
age*household size	0.066(0.019)	0.075(0.017)
NE region	-0.566 (0.056)	-0.405 (0.039)
South region	-0.203 (0.056)	0.055 (0.044)
West region	-0.42(0.065)	-0.124 (0.049)
% Black or non-Black Hisp * %multiunits		0.049 (0.023)
% renter* %multiunits	0.040 (0.017)	0.067 (0.017)