POST-STRATIFICATION CLASSIFICATION AFFECTED BY HOUSEHOLD COVERAGE

Richard Griffin, U.S. Census Bureau, Washington DC 20233

1. Introduction

Post-stratification prior to Dual System Estimation (DSE) is planned for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.). Plans for A.C.E. post-stratification are given in Kostanich (1999). Post-stratification involves grouping sample elements (persons) into approximately "homogeneous" groups with respect to census coverage. Coverage error models for Census data are provided in Wolter (1986). DSE's within post-strata for A.C.E. will use a modified DSE applied to Wolter's model M_{th} (page 341 of Wolter (1986)). This model assumes different capture probabilities in the census and A.C.E. and heterogeneous capture probabilities within census and A.C.E.. The DSE is not consistent for this model and Wolter provides the leading term of the bias. In theory, the DSE is consistent if the capture probabilities are homogeneous in either the census or A.C.E.. This would result in heterogeneous independence and a consistent DSE for model M_{th}. The DSEs for the 1990 Post Enumeration Survey PES 357 post-strata design (see Hogan (1993)) are believed to contain residual heterogeneity. It is not practical to assume that any post-stratification plan we use for A.C.E. will result in homogeneity for either A.C.E. or the census. Thus we will not have heterogeneous independence in any final post-strata and our DSEs will not be consistent. We want to form our post-strata to minimize heterogeneity bias subject to sample size constraints in order to control variance.

ACE is composed of the E-sample and the P-sample. The two samples are based on a common area sample of census blocks. The P sample is designed for the estimation of the number of persons missed by the census and consists of persons enumerated during A.C.E. interviewing. The E sample is designed for estimation of the number of census enumerations that are erroneous and consists of the census enumerations in the common area sample of census blocks.

Richard Griffin is a mathematical statistician in the Decennial Statistical Studies Division of the U.S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Mulry and Spencer (1991) discuss balancing error in DSEs. It is not practical to search the entire census before declaring a P-sample person to be not enumerated or an E-sample enumeration as duplicate or fictitious. Thus the search area for A.C.E. will be limited to one ring of surrounding blocks for selected block clusters. Balancing error occurs due to limiting the search area. Limiting the search area inflates the E-sample estimates of erroneous enumerations and the P-sample estimates of omissions. If these off-setting errors are not in the same proportion, the result is a bias in the DSE referred to as balancing error.

Some of the potential post-strata for A.C.E. are affected by household coverage. For example, one post-strata being considered is "household composition". There are two levels: (1) easy to enumerate and (2) difficult to enumerate. Household composition is a person-level variable. If a household contains more than seven residents, all people withing that household are "difficult". In a single person household, a person is "easy" if he/she is 50 years of age or older. In a two to seven person household, the first two people listed are defined as "easy" if they are both 30 years of age or older and are married. Children of person number one who are less than 18 years of age are "easy" as are parent's of person one who are 50 years of age or older. Children age 18 or older and parent's of person one who are less than 50 years of age are defined as "difficult". "Difficult" household members do not in general cause other household members to be classified as "difficult". There is a concern about a post-stratification variable such as household composition because the results of the household enumeration can affect the post-strata classification of household members.

This paper uses a simple model where the only potential errors are heterogeneity error and balancing error to investigate conditions for which the total error in DSE is decreased by post-stratification by a two category variable affected by the results of household enumeration.

2. General Model Assumptions

1. Assume a closed population of N persons that can be stratified into easy to capture persons and difficult to capture persons. This classification for a given person can be affected by the capture status of other persons in the household.

2. Assume perfect matching within the search area and perfect data collection for captured persons. Sometimes a matching person is classified in the wrong post-strata due to other household members not being captured.

3. Assume duplicates and erroneous enumerations are identified and that there is no nonresponse.

4. With each true strata capture probabilities are homogeneous and the event of being captured in the P-sample is independent of the event of being captured in the E-sample (Causal Independence).

5. Assume the P-sample and the E-sample are 100% of the population. Thus there is no sampling error and we can concentrate on bias.

6. Thus the Peterson model (Wolter's model M_t) holds in each true stratum (homogeneous capture probabilities in both census and A.C.E.). If there were no classification error, the DSE would be consistent.

7. There is no balancing error within each true post-stratum.

8. There are no other errors.

These assumptions indicate that if there were no classification errors the DSE in each true post-strata would be essentially unbiased for a large population. However, after data collection we post-stratify both the P-sample and E-sample enumerations based on the persons captured in the household. Thus some P-sample persons and some E-sample persons are classified in the wrong post-strata. As a result of these classification errors, balancing error and heterogeneity error are present in the observed post-strata. The "true" post-strata classification is based on all members of the household and the observed post-strata classification is based on the persons actually captured for the P-sample or E-sample. Causal independence still holds for the observed post-strata.

3. Balancing error

As in Mulry and Spencer (1991), for a given post-strata let b denote the proportion of correct census enumerations that are in the P-sample search area and let g denote the ratio of the number of correct census enumerations that are in their E-sample search area to the number that are in their P-sample search area. Thus bg is the proportion of correct census enumerations that are in their E-sample search area. For each true post-strata we assume that g = 1.

The population sizes and capture probabilities for a given post-strata are given in Table 1. The subscript + indicates summation over the possible values of the subscript and population sizes in parenthesis are not observed.

Table 1 - Capture Probabilities and Population sizes in Post-strata

		Census Capture Probability/Population Size		
		In	Out	Total
ACE	In	P_{11} / N_{11}	P_{12} / N_{12}	P_{1+} / N_{1+}
	Out	P_{21} / N_{21}	$P_{22} / (N_{22})$	$P_{2+} / (N_{1+})$
	Total	P_{+1} / N_{+1}	$P_{+2} / (N_{+1})$	$P_{++} / (N_{++})$

In Table 1, $N_{++} = N$ is the total size of the population. The census count is not the same as N_{+1} because erroneous enumerations are not included in N_{+1}. Even if one could observe the population sizes in the first row and first column, the counts in parenthesis would not be observed and would have to be estimated.

The target DSE is obtained by dividing the number of people enumerated in the census (corrected for erroneous enumerations and duplicates), N_{+1}, by an estimate from ACE of the proportion of the population that were enumerated in the census. The resulting target DSE is given by, $N^T = (N_{+1})(N_{1+})/N_{11}$.

We do not observe N_{+1} and N_{11} due to limiting the search area. The observed values are as follows:

$$N_{11}^O = bN_{11}$$

and

$$N_{\cdot 1}^O = bgN_{\cdot 1}$$

Thus the observed DSE is given by

$$N^O = \frac{N_{1\cdot} N_{\cdot 1}^O}{N_{11}^O}$$

For the true post-strata g = 1 and $N^O = N^T$. For observed post-strata we have the balancing error given by

$$N^O - N^T = \frac{(g-1)N_{\cdot 1}N_{1\cdot}}{N_{11}}$$

4. Heterogeneity Bias

From Wolter (1986) for model M_{th} the bias due to heterogeneity is given by

$$N^T - N = Bias(N^T) = \frac{-N\sigma(P_{1\cdot}, P_{\cdot 1})}{\sigma(P_{1\cdot}, P_{\cdot 1}) + \overline{P_{1\cdot}}\ \overline{P_{\cdot 1}}}$$

To define the new terms, let the subscript i denote person i and j denote person j so that for example P_{i+1} is the census capture probability for person i and P_{j+1} is the census capture probability for person $j \neq i$. Due to heterogeneity in the observed post-strata these capture probabilities may not be equal. Then

$$\sigma(P_{1\cdot}, P_{\cdot 1}) = N^{-1}\sum_i (P_{i1\cdot} - \overline{P_{1\cdot}})(P_{i\cdot 1} - \overline{P_{\cdot 1}})$$

$$\overline{P_{1\cdot}} = N^{-1}\sum_i P_{i1\cdot}$$

$$\overline{P_{\cdot 1}} = N^{-1}\sum_i P_{i\cdot 1}$$

5. Total Error Decomposition

Since we are assuming that the only errors in the observed post-strata are balancing error and heterogeneity bias, the total error is given by the sum of the balancing error and the heterogeneity bias (total error = balancing error + heterogeneity bias) or

$$N_O - N = (N^O - N^T) + (N^T - N)$$

6. Application

Balancing error

Denote the "easy" observed post-strata by 1 and the "difficult " post-strata by 2 and assume equal sample sizes of persons in the E and P samples for the true post-strata. Since we are assuming no sampling the sample size for post-strata k is $N_{k+1} = N_{k1+}$.

Denote the sample size for post-strata i by n_i and let R_{sij} denote the "switching" rate for sample S (E or P), from post-strata i to post-strata j. Thus R_{E12} is the rate at which E sample persons in true post-strata 1 are observed in post-strata 2 and R_{P21} is the rate at which P sample persons in true post-strata 2 are observed in post-strata 1. Recall the definitions of g and b in section 3 and assume for the true post-strata that $g_1b_1 = g_2b_2 = z$. Here we are assuming that the proportion of correct enumerations that are in the E-sample search area is the same in the "easy" and "difficult" post-strata. Thus the number of correct enumerations in true post-strata k is given by zN_{k+1}.
Thus for the observed post-strata

$$g_1 = \frac{N_{1\cdot 1}(1 - R_{E12}) + N_{2\cdot 1}R_{E21}}{N_{1\cdot 1}(1 - R_{P12}) + N_{2\cdot 1}R_{P21}}$$

$$g_2 = \frac{N_{2\cdot 1}(1 - R_{E21}) + N_{1\cdot 1}R_{E12}}{N_{2\cdot 1}(1 - R_{P21}) + N_{1\cdot 1}R_{P12}}$$

Note that if $R_{E12} = R_{P12}$ and $R_{E21} = R_{P21}$ then $g_1 = g_2 = 1$. Thus if the proportion of true "easy" post-strata E sample persons who are observed in the "difficult" post-strata is not equal to the same proportion for the P sample or if the proportion of true "difficult" post-strata E sample persons who are observed in the "easy" post-strata is not equal to the same proportion for the P sample, then there will be balancing error in the observed post-strata. For example, one could conjecture that the A.C.E. interview will do a better job of capturing all members of each household so that R_{E21} will be greater than R_{P21}. The opposite could also be true.

Heterogeneity Bias

Let r_{ij} denote the proportion of the total population N_i in true post-stratum i that are observed in post-stratum j. Assume that $r_{ij} = .5(R_{Eij}+R_{Pij})$. For each observed post-strata there are only two possible capture probabilities. For observed post-strata 1, $N_1(1-r_{12})$ have capture probabilities P_{1+1} and P_{11+} while N_2r_{21} have capture probabilities P_{2+1} and P_{21+}. For observed post-strata 2, $N_2(1-r_{21})$ have capture probabilities P_{2+1} and P_{21+} while N_1r_{12} have capture probabilities P_{1+1} and P_{11+}. Given this the heterogeneity bias for each observed post-strata is easily computed.

Detailed example

First consider a true easy post-strata with a capture probability of 0.9 for each person in the P-sample and E-sample, there are 1,000 persons in the target population. Assuming causal independence it is a simple matter to distribute these 1000 persons into the categories defined in Table 1. Next consider true difficult post-strata of 1000 persons with a capture probability in the P-sample and E-sample of 0.45. The ratio of the true easy post-stratum capture probability to the true difficult post-strata capture probability is 2.

There are 6 input parameters for this example. The first is the post-strata sizes which we will not vary and fix at 1,000 in each true post-strata. The other 5 are as follows:

1) Re21=.05 = the proportion of true E sample persons in post-stratum 2 (difficult) who are observed in post-stratum 1(easy);

2) Re12/Re21= .2 = the ratio of true E sample persons in post-stratum 1 who are observed in post-stratum 2 to Re21;

3) Rp/Re= .5 = the ratio of switching from post-strata for the P sample to switching post-strata for the E sample;

4) EasyP/DiffP = 2 = the ratio of the capture probability in true post-stratum 1 to true post-stratum 2;

and 5) Easy P = .9 = the capture probability in true post-stratum 1.

Under these conditions, there will be 1030 persons in the observed easy post-strata and about 3.6% of these persons are in the true difficult post-strata. There will be 970 persons in the observed difficult post-strata and about 0.8% of these persons will be in the true easy post-strata. For this example, the balancing bias, heterogeneity bias, and total error are computed as described in sections 3, 4 and 5 respectively. The net bias is the sum of the total bias in observed post-strata 1 and 2. The resulting relative net bias is -0.01181. For comparison purposes we consider a post-stratum formed by combining the easy and difficult post-strata into one post-stratum. This post-stratum would have no balancing bias but substantial heterogeneity bias. The resulting relative bias is for this example is -0.1. The ratio of the net relative bias when the two post-strata are formed to the net relative bias when they are combined in this example is about 0.12 indicating that the increased balancing error due to forming the two post-strata is more than compensated for by decreased heterogeneity bias.

In this example it is much better to form the 2 post-strata due to a high model bias for the combined post-strata and a relatively low balancing bias over the 2 post-strata. This is caused by a capture probability ratio (easy/difficult) of 2. If Re21 = .5 and the capture probability ratio is reduced to 1.15, the combined model bias is significantly reduced relative to the balancing bias over the 2 post-strata and the ratio of the net bias from 2 post-strata to the model bias from combining is -1.33 indicating it is better not to form 2 post-strata.

To further analyze the relationship between balancing bias and heterogeneity, fix the capture probability in the easy post-strata at 0 .9 and fix the ratio of E sample switching to the difficult post-strata to E sample switching to the easy post-strata (Re12/Re21) at 0 .2. Now for a given Re21, as Rp/Re decreases there is less switching for the P sample (relative to the E sample) and increasing balancing bias if 2 post-strata are formed. For the combined post-strata as EasyP/DiffP increases there is more model bias due to heterogeneity. It is better to combine when there is not too much heterogeneity for a given amount of balancing bias if you form 2 post-strata. If the balancing bias increases the amount of heterogeneity you can have and still be better to combine increases. Thus, for a given Re21 value as Rp/Re decreases there is more balancing bias and the value of EasyP/DiffP at which it becomes better to combine increases. Table 2 shows results for various combinations of Re21 and Rp/Re. For example, for Re21 = .5 and Rp/Re = .25 the EasyP/DiffP ratio at which it becomes better to combine is 1.22. For ratios higher than this it is better to form 2 post-strata since there would be to much

heterogeneity in the combined post-strata. If Re21 remains at .5 but Rp/Re decreases to .1, there is more balancing bias in forming 2 post-strata and the EasyP/DiffP ratio at which it is better to combine increases to 1.26. More heterogeneity is allowed in the combined post-strata due to the increased balancing bias if 2 post-strata are formed. As shown in Table 2 as Re21 decreases the levels of heterogeneity at which it is better to combine decreases. This is because the overall level of balancing bias if 2 post-strata are formed decreases so gains from decreased heterogeneity are not offset by balancing bias. For small values of Re21, it is almost always better to form 2 post-strata to decrease heterogeneity bias.

7. Conclusion

If a potential two level post-stratum affected by household coverage has capture probabilities that are "significantly" different (ratio higher or lower than 1), it is likely that the reduction in heterogeneity bias by using the post-stratum is not offset by the balancing bias caused by different proportions of the P and E samples switching post-stratum level due to household coverage. It is likely that a capture probability ratio of 1.3 (or 1/1.3 = 0.769) or higher (or lower) indicates a strong candidate for selection as a final post-stratum in spite of balancing bias. Logistic regression modeling has been used to isolate independent variables that are strong predictors of census capture probability. Odds ratios are part of the output of logistic regression modeling. A odds ratio is equal to the ratio of odds for census capture probabilities at the two levels of the potential post-stratifier. Thus the results of logistic regression together with the analysis in this paper have been useful in selecting post-strata for Census 2000. Primarily due to concerns about balancing error, given that other variables can be used to group persons into homogeneous post-strata, no variables affected by household coverage will be used for Census 2000 A.C.E. post-stratification. Research will continue for 2010.

References

Hogan, H. (1993), "The 1990 Post-Enumeration Survey: Operations and Results", *Journal of the American Statistical Association*, 88, pp.1047-1060.

Kostanich, D., Griffin R., and Fenstermaker, D.(1999), "Accuracy and Coverage Evaluation Survey Plans for Census 2000", document prepared for the March 19,1999 meeting of the National Academy of Science Panel to Review the 2000 Census.

Mulry M. And Spencer B. (1991), "Total Error in PES Estimates of Population", *Journal of the American Statistical Association*, 86, pp.839-863.

Wolter, K. (1986), "Some Coverage Error Models for Census Data", *Journal of the American Statistical Association*, 81, pp.338-346.

TABLE 2: CAPTURE PROBABILITY RATIO: COMBINE OR FORM 2 POST-STRATA

Example: For Re21 = .4 and Rp/Re = .1, the largest capture probability ratio indicating the one combined post-stratum should be formed is 1.21. For ratios larger than 1.21, form 2 post-strata.

Re21	Rp/Re	EasyP/DiffP
.5	.5	1.15
.5	.4	1.18
.5	.25	1.22
.5	.1	1.26
.5	.05	1.27
.4	.5	1.12
.4	.4	1.14
.4	.25	1.17
.4	.1	1.21
.4	.05	1.22
.25	.5	1.07
.25	.4	1.08
.25	.25	1.11
.25	.1	1.13
.25	.05	1.13
.1	.5	1.02
.1	.4	1.03
.1	.25	1.04
.1	.1	1.05
.1	.05	1.05
.05	.5	1.01
.05	.4	1.01
.05	.25	1.02
.05	.1	1.02
.05	.05	1.02