ALTERNATIVE OVERSAMPLING OPTIONS FOR LOW MAIL RESPONSE AREAS IN THE AMERICAN COMMUNITY SURVEY[1]

Anthony Tersine and Mark Asiala, U.S. Census Bureau
Anthony Tersine, U.S. Census Bureau, Demographic Statistical Methods Division, Washington, D.C. 20233

KEY WORDS: American Community Survey, sample design, oversampling

Abstract

The Census Bureau's American Community Survey (ACS) is planned to be an annual sample of three million housing units mailed out in monthly panels. Data collection for each monthly panel extends over a three-month period: mailout/mailback in the first month, telephone follow-up in the second month for addresses where a telephone number can be obtained, and personal-visit followup in the third month for a one-third subsample of the remaining nonrespondents. Areas with low mail response will have a larger percentage of cases going to personal-visit, and thus will have a larger variance estimate because of the subsampling in this phase. This paper will examine several options of oversampling these areas to increase the reliability of the estimates.

1.0 Introduction

The ACS is designed as a monthly mail-out survey with follow-up by Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) operations during a three month interview cycle. The ACS is an annual survey of three million addresses with one-twelfth of the sample mailed out each month. All households with a mailable address are sent a mail questionnaire during the first month of the interviewing cycle. During the second month, all households which did not return a mail form and for which we can obtain a telephone number are sent to CATI. During the third month, all households which did not return a mail form or for which we did not obtain a CATI interview are sent to CAPI. Those eligible for CAPI are sub-sampled at two different rates: 2-in-3 for units without a mailable address and 1-in-3 for all other units.

This paper focuses on the last component. One of the objectives of the ACS is to produce reliable tract level estimates. Currently, tracts (or any other geographical unit) with low mail response rates tend to have a larger percentage of their total sample represented by the units which were sub-sampled at the 1-in-3 rate in the personal-visit component. This inflates the variance and hence the coefficient of variation (CV) for these tracts. This paper presents the findings of our research into methods to make the CVs more equal across all tracts. The main component of this research is increasing the CAPI rates in low mail response areas and decreasing the overall sampling rate in high response areas. Since the ACS will only produce tract level estimates based on a 5-year average, the results presented here are 5-year average tract level estimates.

2.0 Current Sample Design

The current sample design involves four sampling rates based on the number of housing units in the governmental unit and the census tract. We define a governmental unit to be a county, school district, American Indian area (including Alaska Native Areas and Hawaiian Homelands), functioning place, or functioning Minor Civil Division (only in twelve states). If a housing unit falls in more than one governmental unit, we use the number of housing units for the smallest governmental unit. The sampling rates are as follows:

1. In a governmental unit with less than 800 housing units => 3 * base rate
2. In a governmental unit with 800 or more housing units but less than or equal to 1200 housing units => 1.5 * base rate
3. In a tract with more than 2000 housing units => 0.75 * base rate
4. Not in 1., 2., or 3. above => base rate

The base rate is determined to give three million addresses a year. For the 2003 ACS, the base rate is approximately 2.5 percent.

[1]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.

3.0 Methodology

The basis for our methodology is to identify low mail response rate tracts along with high response rate tracts which we would expect in the ACS. As a predictor for this data, we used the Census 2000 Supplementary Survey (C2SS) and the 2000 ACS response rates where possible. In tracts which were not in sample in 2000 for ACS, we predicted the ACS response rate based on preliminary Census 2000 long form response rates using a logit model. CATI and CAPI workload and interview rates were also modeled based on the 2000 data.

Using this data and model parameters along with per case cost estimates, we constructed different cost neutral scenarios. Cost neutral means that the total cost of all three data collection modes is the same, but that the cost of each data collection mode can vary. They all involved setting three thresholds which divided the tracts into four categories:

1. 2-in-3 CAPI subsampling
2. 1-in-2 CAPI subsampling
3. 1-in-3 CAPI subsampling (status quo)
4. 1-in-3 CAPI subsampling with reduction in the overall sample

The reduction in sample in the fourth category is the offset which covers the additional expense of the oversampling in the first and second categories.

We then calculated CVs for a 10% characteristic for all of the tracts. These CVs are summarized for the categories along with different breakdowns based on the governmental unit sampling rates.

3.1 Models

Since the C2SS and the ACS were not conducted in all counties, we needed to model the data for those areas not in sample. When this research gets implemented, we will be able to use the actual ACS data. Using the response rates from Census 2000 and Census 2000 Supplementary Survey / ACS mail, CATI, and CAPI response rates, response rates and workloads are assigned to each tract to calculate the variance for each tract. Our variance calculation stratifies (see Section 3.2) on CAPI / non-CAPI mode so these response rates will help us determine the population and sample estimates for each stratum and the number of cases in each mode for cost calculations. The expected ACS mail, CATI and CAPI response rates are modeled as follows:

1. ACS non-mailable rate = average tract non-mailable rate in each sampling stratum

2. ACS mail response rate = exp(-2.41831 + 3.60512 * Census 2000 long form mail response rate) / [1 + exp(-2.41831 + 3.60512 * Census 2000 long form mail response rate)] (R-squared = 0.3174) (by tract)

3. ACS CATI workload rate = 6.74766 + 0.39772 * ACS mail non-interview (R-squared = 0.2712) (by county)

4. ACS CATI interview rate = -0.65316 + 0.33337 * ACS CATI workload rate (R-squared = 0.5129) (by county)

5. ACS CATI Late Mail Return rate = 0.15394 + 0.17353 * ACS CATI workload rate (R-squared = 0.1483) (by tract)

6. ACS CAPI interview rate = -1.75339 + 0.91006 * ACS CAPI workload rate (R-squared = 0.9491) (by county)

In all cases, if the tract was in the Census 2000 Supplementary Survey / ACS then the Census 2000 Supplementary Survey / ACS data was used. Tracts which were not in the Census 2000 Supplementary Survey / ACS used the modeled rates.

3.2 Variance Calculations

In order to evaluate the different oversampling options, the coefficient of variation for a 10% characteristic was calculated. We estimated the variance using a basic stratified binomial distribution described below.

The proportion P of persons with characteristic X over all sampling strata i is estimated as follows:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{h} N_i \hat{p}_i$$

The variance of $\hat{p}$ can be estimated as follows:

$$\hat{V}(\hat{p}) = \frac{1}{N^2}\sum_{i=1}^{h} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{p}_i \hat{q}_i}{n_i}\right)$$

where h is the number of sampling strata. In our case, there will be eight possible strata at the tract level: four for the initial governmental unit sampling rates crossed with two for CAPI / Non-CAPI respondents.

The CV would then be defined as:

$$CV(\hat{p}) = \frac{\sqrt{\hat{V}(\hat{p})}}{\hat{p}}$$

3.3 Oversampling

The oversampling was implemented at two different levels. The tracts were ranked according to predicted mail response rate. Three percentile thresholds were designated. All tracts which were in a percentile less than or equal to the first threshold had their mail/CATI non-interviews sub-sampled at a 2-in-3 rate. All tracts which were in a percentile greater than the first threshold but less than or equal to the second threshold used a 1-in-2 CAPI subsampling rate. All tracts in a percentile greater than the second threshold but less than or equal to the third threshold used the default 1-in-3 CAPI subsampling rate. All tracts in a percentile greater than the third threshold used the default 1-in-3 CAPI subsampling rate and also had their initial sample reduced by a fixed ratio. This ratio was calculated to reduce the cost of the non-oversampled tracts to offset the increased costs of the oversampling thus making this a cost neutral implementation. Different oversampling scenarios were examined. Each scenario uses different thresholds for determining which tracts are over-sampled at the higher 2-in-3 or 1-in-2 CAPI subsampling rates.

4.0 Results

We looked at several scenarios in our research and present some here. Table 1 below gives the percentile and response rate cutoffs for five scenarios. Scenarios 1–4 are possible scenarios to implement, and scenario 5 is an example of reducing the mailout (i.e. a large amount of oversampling).

Thus for scenario 1, all tracts whose predicted mail response rate is in the 0–7th percentiles (less than or equal to 16.7% response rate) will have a CAPI subsampling rate of 2-in-3. All tracts whose predicted mail response rate is in the 7–16th percentiles (greater than 16.7% but less than or equal to 27.3% response rate) will have a CAPI subsampling rate of 1-in-2. Tracts in the 16–50th percentiles (greater than 27.3% but less than or equal to 46.0%) have no change made to either their CAPI subsampling rate or their overall sample. Finally, tracts in the 50th percentile and above (predicted mail response rate of greater than 46.0%) retain the 1-in-3 CAPI subsampling but have only 80.2% of there initial overall sample. Note that only this category has its sample reduced.

What effect does the oversampling described in Table 1 have on the CVs? Table 2 presents the changes to the CVs based on the oversampling for scenarios 1 and 5. Table 3 presents the changes to the CVs based on the oversampling for scenario 1 crossed by the governmental unit sampling rates.

In Table 2, scenario 1, the median CV for the 2-in-3 CAPI cases is reduced from 25.4% to 18.1%, but for the cases where we reduced the initial sample, the median CV increased from 17.1% to 19.0%. In this scenario, the CVs for the lowest response cases are now better than for the highest response cases. This means we are probably doing too much oversampling. We want to improve the CVs in the low response areas, which does happen, but not at the expense of making the CVs for the high response areas worse than the low response areas. We observe the same changes in CVs for scenario 5, but with a much larger magnitude. We observed the same patterns in scenarios 2-4 with respect to the CV. These results indicate we are doing too much oversampling, and we need to look at oversampling a smaller proportion of the cases. Another option in this scenario is to use several cutoffs for reducing the initial sample size.

In Table 3, we observe, for the most part, the same patterns seen in Table 2. That is, the median CVs for the lowest response cases are now better than for the highest response cases. The exception to this is for the column '3 * Base Rate' where the median CV for the highest response areas is better than for the lowest response areas. We observed the same patterns in scenarios 2–5 with respect to the CV.

Table 1. Cutoffs for Oversampling

Scenario	Upper Limit of 2-in-3 sampling		Upper Limit of 1-in-2 sampling		Upper Limit of Standard 1-in-3 sampling		Percent Of Initial Sample Remaining in Reduced 1-in-3 sampling
	Percentile	Response Rate (%)	Percentile	Response Rate (%)	Percentile	Response Rate (%)	
1	7	16.7	16	27.3	50	46.0	80.2
2	8	18.8	16	27.3	51	46.4	78.9
3	9	20.0	14	25.0	50	46.0	80.6
4	10	21.1	14	25.0	50	46.0	79.3
5	40	41.6	56	49.9	56	49.9	15.6

Table 2. CVs for Oversampling Options (Tract-level CVs for a 5 year estimate of a 10% characteristic).

Scenario	New CAPI Rate	Original/New Sample Design	Number of Tracts[1]	Mean	Minimum	Q1	Median	Q3	Maximum
1	All	Original	64,285	19.78	5.74	16.06	18.46	21.82	92.29
		New	64,335	19.83	6.73	16.52	18.95	21.83	89.84
	2-in-3	Original	5,023	28.84	9.32	21.45	25.39	32.25	92.29
		New	5,081	21.06	8.04	15.39	18.07	22.92	89.84
	1-in-2	Original	5,604	23.01	9.46	18.77	21.58	25.88	70.27
		New	5,610	19.24	8.00	15.65	17.89	21.47	72.05
	1-in-3 (without reduction)	Original	21,290	20.07	6.78	16.98	19.17	22.08	64.71
		New	21,290	20.07	6.78	16.98	19.17	22.08	64.71
	1-in-3 (with reduction)	Original	32,368	17.63	5.74	14.95	17.07	19.39	89.70
		New	32,354	19.58	6.73	16.58	19.03	21.64	89.69
5	All	Original	64,285	19.78	5.74	16.06	18.46	21.82	92.29
		New	64,044	23.59	5.70	15.24	21.56	29.28	93.81
	2-in-3	Original	25,376	22.80	7.42	18.06	20.91	25.17	92.29
		New	25,442	17.14	5.70	13.73	15.62	18.61	89.84
	1-in-2	Original	9,864	18.36	6.78	16.05	17.97	20.18	50.39
		New	9,864	15.86	5.78	13.93	15.50	17.30	50.39
	1-in-3 (with reduction)	Original	29,045	17.63	5.74	14.89	17.02	19.41	89.70
		New	28,738	31.96	12.15	26.27	29.61	35.74	93.81

[1] The difference in the number of tracts between the original and new is due to a requirement on the tract interview sample size (at least five housing units) for the 5 years.

Table 3. Median CVs for Oversampling (Scenario 1) by Governmental Unit Sampling Rate (Tract-level CVs for a 5 year estimate of a 10% characteristic).

Scenario	New CAPI Rate	Original/New Sample Design	All	3 * Base Rate	1.5 * Base Rate	Base Rate	0.75 * Base Rate
1	All	Original	18.46	12.86	16.37	20.10	17.08
		New	18.95	13.67	17.03	20.24	17.86
	2-in-3	Original	25.39	20.98	20.81	26.33	20.30
		New	18.07	14.29	15.82	18.88	14.61
	1-in-2	Original	21.58	16.17	18.84	23.33	19.35
		New	17.89	13.68	15.95	19.29	16.09
	1-in-3 (without reduction)	Original	19.17	14.15	17.33	20.77	17.75
		New	19.17	14.15	17.33	20.77	17.75
	1-in-3 (with reduction)	Original	17.07	11.98	15.27	18.18	16.32
		New	19.03	13.41	17.01	20.25	18.29

5.0 Future Research

The other option we have yet to explore is basing the cutoffs on the combined mail and telephone response rates. This option is possible since the costs for mail and telephone are about the same, and there is no subsampling of housing units in the telephone phase. Looking at the combined mail and CATI response rates and trying to balance the oversampling is the next step in our research.

6.0 Conclusions

Oversampling in the above method provides some benefits over the default method. In particular, the CVs for the tracts with the lowest predicted mail response rates improve with only a small impact on the higher mail response tracts. It is possible, therefore, to obtain more equal CVs across all tracts using oversampling in the manner outlined. Based on the different scenarios examined, we need to do further research in oversampling options to reduce the amount of oversampling. The results seem promising given the objective of the ACS, but more work is needed to decide the exact scenario.