POST-STRATIFICATION FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALAUATION SURVEY

Dawn E. Haines and Eric Schindler
Eric Schindler, U. S. Census Bureau, Washington, DC 20233

Keywords: Estimation; Logistic Regression; Error Estimation

Abstract. The final estimates of population for the 1990 Census Post Enumeration Survey (PES) used the post-stratification variables race/Hispanic origin, age/sex, tenure, Census Region, and size of urban area. For the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Survey, these variables and others were considered. Significant variables were identified using logistic regression modeling on the 1990 PES data. Several sets of target estimates were developed using all of the significant variables (and some of their interactions) and demographic analysis. Dual system estimates were simulated for a number of post-stratification models with subsets of the significant variables. Estimates of population, variance, and bias were calculated for the 1990 PES post-stratum groups, states, congressional districts, and a selected set of large cities. The variance was also adjusted to reflect the increased sample size and design of the A.C.E. survey. These results are one component of the decision-making process for selecting the final A.C.E. post-stratification variables.

Introduction

Dual system estimation for the 1990 PES was first implemented with 1392 post-strata with 116 post-stratum groups reflecting the sample design and twelve age/sex categories within each group. Some of the post-strata were very small, requiring that the variance estimates be smoothed. The smoothing process was not well understood. This contributed to the Secretary of Commerce's decision not to adjust the official census estimates to correct for the estimated undercount.

After the Secretary's decision the Census Bureau developed a revised post-stratification design with 357 post-strata. (See Hogan, 1993) This design eliminated the smallest post-strata of the original design, calculated jackknife variance estimates, and has generally been accepted as producing the best available estimates from the 1990 PES data. The 357 post-strata are defined by the variables race/Hispanic origin, age/sex, tenure, Census region, and size of urbanized area.

For several years the Census Bureau has been testing additional variables for post-stratification for the Census 2000 A.C.E.. The objective is to produce a design which does as well as or better than the PES design for most purposes and can be implemented in a shorter time frame. The 1990 PES data has been used for this research. Logistic regression analysis was used to identify several additional variables significantly correlated with coverage in the census. (See Haines and Hill, 1998.) Dual system estimates were calculated for several post-stratification models to determine the effect on estimates at a number of levels. (See Schindler, 1999.) Some variables were dropped because their practical impact on the estimates was negligible; others because of problems developing consistent definitions or for operational considerations.

The Census Bureau has selected a post-stratification design for the Census 2000 A.C.E. survey which retains the race/origin, age/sex, tenure, and region variables from the 1993 post-stratification of the 1990 PES. A size of consolidated metropolitan statistical area (CMSA) variable replaces the 1990 PES size of urbanized area variable which will not be available in time for A.C.E. estimation. Two new variables for Census 2000 are the tract level return rate which indicates the level of neighborhood cooperation in the census and the type of enumeration area (TEA) which indicates how the census delivers and collects the form. Since large metropolitan areas are primarily collected by the mailout/mailback (MO/MB) procedure, the TEA variable will be crossed by the CMSA variable. These new variables are expected to help define post-strata which are homogeneous with respect to coverage in the census, reducing the heterogeneity bias which occurs when disparate groups are kept together in dual system estimation. Post-strata will also have larger sample sizes than in 1990 in order to control variances which were high for many post-strata even for the final PES design. The Census 2000 A.C.E. post-stratification design has 448 post-strata, shown in Appendix A, before final collapsing to ensure acceptable sample sizes.

The remaining sections of this paper discuss the A.C.E. post-stratification model, simulations of the A.C.E. post-stratification with the 1990 PES data, and

---

Dawn Haines and Eric Schindler are mathematical statisticians in the Decennial Statistical Studies Division of the US 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.

some brief conclusions.

The A.C.E. Post-stratification Model

THE SAMPLE

The A.C.E. sample consists of about 300,000 housing units in 11,000 block clusters, approximately twice the size of the 1990 PES. Sample size increases are not uniform. In 1990, inner city Black and Hispanic groups were oversampled by a large factor. The A.C.E. also oversamples these groups but not as much. The A.C.E. sample size for Blacks should be about 20% larger than the PES sample and the Hispanic sample should be about 60% larger. Three improvements in the A.C.E. sample design which should control sampling error are:

- The decreased oversampling of inner city minority areas leaves more sample for other areas and reduces weight variation.
- Block clusters with potential coverage problems have been oversampled for the A.C.E., reducing their impact on the estimates.
- More small block clusters have been selected.

VARIABLES

Simulations using the 1990 PES data have been run for various post-stratification designs over the last two years. Statistical significance was determined for a number of variables using logistic regression methods (Haines and Hill, 1998). Some potentially powerful post-stratification variables, such as whether a housing unit could be associated with a mail return or which household members were part of a nuclear family were dropped because they could not be uniformly defined. Other less significant variables, such as Census Region, survived the cut for non-Hispanic White and Some other race owners, about half the population, since it did not adversely affect variances.

The following variables will appear in the Census 2000 A.C.E. post-stratification:

- Race/Hispanic origin:
  The seven domains are approximately: (1) American Indians or Alaska Natives (AI/AN) living on Reservations, (2) AI/AN not living on Reservations,(3) Hispanic, (4) Non-Hispanic Black, (5) Native Hawaiian or Pacific Islander (NH/PI), (6) Non-Hispanic Asian, and (7) Non-Hispanic White or "Some other race." Persons are usually assigned to the lowest numbered domain indicated.
- Age/Sex:
  Seven categories: (1) 0 - 17, (2) 18-29 Male, (3) 18-29 Female. (4) 30-49 Male, (5) 30-49 Female, (6) 50 and over Male, and (7) 50 and over Female
- Tenure:
  Two categories: (1) Owner and (2) Non-owner
- Census Region
  Northeast: Maine through Pennsylvania
  Midwest: Ohio through Kansas / North Dakota
  South: Delaware through Oklahoma / Texas
  West: New Mexico / Montana and west
- CMSA/TEA:
  The urbanized area size class variable used in the 1990 PES design is being replaced by the Census 2000 total population count in the Consolidated Metropolitan Statistical Area (CMSA), as defined in 1999. This variable is crossed by the Type of Enumeration area to define four categories. Housing units collected by other than the MO/MB methods, mostly Update/Leave/Mailback (U/L), will be placed in the fourth category to capture differences in quality of the Master Address File depending on whether the address was taken from the U.S. Postal Service and census forms were mailed (mostly urban areas) or it was created or updated by Census Bureau staff who hand-delivered the census forms (mostly rural areas). The other three categories are: (1) MO/MB in the largest 10 CMSAs, (2) MO/MB in other CMSAs or MSAs with unadjusted census counts of 500,000 or more persons, and (3) MO/MB in small MSAs and non-MSA areas. The first two groups cover about 30% of the population. About 20% of the population are in MO/MB areas of small MSAs or non-MSA areas, and the last 20% are in non-MO/MB areas.
- Tract Return Rate:
  Two categories: For the Census 2000 A.C.E. six cutoffs for this variable will be defined separately for the six groups Non-Hispanic White owners, Non-Hispanic White non-owners, Non-Hispanic Black owners, Non-Hispanic Black non-owners, Hispanic owners, and Hispanic non-owners such that three quarters of each of these groups will be classified as living in high return tracts for the group, which should correlate to "easy-to-count" and the rest in low return, presumably "hard-to-count," tracts. A Hispanic owner and a Hispanic non-owner living in the same tract may be classified as living in a low return rate tract for Hispanic owners and as living in a high return rate tract for Hispanic non-owners, respectively.

COLLAPSING OF POST-STRATA

The 357 post-stratum design for the PES was based on the variables race/origin (five categories), age/sex, tenure, region, and urbanized area size (three categories).

These variables define 840 post-strata, many of which had very small sample sizes. In order to obtain sufficient sample sizes, region was combined for Blacks and Hispanics in the small urbanized areas and in the nonurbanized areas, all geographic indicators were combined for Asians and Pacific Islanders, and all variables except age/sex were combined for AI/ANs living on reservations. Even with this collapsing, there are nine post-strata with sample sizes with less than 100 persons in the independent sample which have the expected high variances.

The post-stratification variables for the A.C.E. define 3136 cells so substantial collapsing based on the expected A.C.E. sample sizes is required. Most of this collapsing, shown in Appendix A, is being specified in advance. It is possible to maintain all of the variables only for the Non-Hispanic White owners. There are fewer Non-Hispanic White non-owners, so region which is the weakest variable as an indicator of coverage differences, has been dropped. There are even fewer Non-Hispanic Blacks and Hispanics, so the large and medium CMSA/MSA MO/MB groups are combined as are the small and non-MSA and non-MO/MB groups. There are even fewer persons in the other four race/origin groups, and only tenure and age/sex are not being collapsed.

Additional collapsing of the 448 remaining post-strata will occur during the estimation process for post-strata with less than 100 persons in the independent sample. The NH/PIs and the AI/ANs not living on reservations post-strata are the most likely candidates for additional post-stratum collapsing.

Simulations: A.C.E. vs. PES

The 1990 PES used and the 2000 A.C.E. will use the Dual System Estimation (DSE) methodology (Hogan, 1993). A sample, called the Person or P sample, of clusters of blocks (5,200 in 1990; 11,000 in 2000) with about 30 housing units each, is selected and independently reenumerated. The extract of data defined persons from the census for the same block clusters is called the Enumeration or E sample. The two samples and census data from a sample of surrounding areas are compared in order to determine which E sample persons were correctly enumerated and which P sample persons can be matched to a census enumeration. For each post-stratum, the DSE is given by:

$$DSE = (C - II) \times \frac{E - EE}{N_E} \times \frac{N_P}{M} \quad \text{where:}$$

- the first term, slightly less than the census count, adjusts for persons in the census (C) with insufficient information (II, most or all person data imputed, persons are not data defined) who cannot be possibly matched to P sample persons,
- the second term, slightly less than 1, adjusts for persons listed in the E sample in A.C.E. block clusters who were erroneously enumerated because either they should not have been enumerated or they were enumerated in the wrong place or they were enumerated more than once or some critical data for matching, such as name, is missing (E-EE), and
- the third term, slightly greater than 1, adjusts for persons in the P Sample who were matched, and presumably collected, in the census (M).

Coverage correction factors (CCF) for a post-stratum are obtained by dividing the DSE by the census count. Synthetic estimates for any subpopulation are calculated by adding the products of the post-stratum CCFs and the subpopulation post-stratum census counts.

For the simulations that follow, two influential block clusters were dropped from the 5180 block clusters with persons in the 1990 PES. Standard errors were estimated using a simple jackknife procedure dropping out one of the 1990 PES block clusters at a time. The actual A.C.E. standard errors, which will be calculated by a stratified jackknife procedure, should be smaller because of the larger A.C.E. sample size and the more equal weights of the Census 2000 A.C.E. sample design. Also, the definitions of the race/Hispanic origin groups in the simulations follow the 1990 PES definitions which are slightly different than the A.C.E. definitions.

Table 1 shows the effect of adding the last three post-stratification variables to an initial model with only race/origin, age/sex, and tenure. Adding the MSA/TEA variable increases the estimates slightly, but adding region for Non-Hispanic White owners decreases them slightly. Finally, the tract level return rate variable adds to the estimates. All the A.C.E. estimates are within 0.05 percent of each other and the 1990 PES estimate.

Table 1: Estimates for 5 Post-stratification Models

Model	Estimate	CCF	StdError
Census	242,012,129	1.000000	n/a
R/O A/S T	246,151,720	1.017105	0.001823
" + msa/tea	246,170,459	1.017182	0.001828
" + region	246,112,202	1.016942	0.001808
" + return	246,225,017	1.017408	0.001799
1990 PES	246,194,742	1.017283	0.001824

We expect the CCFs for minorities to be higher than those for non-Hispanic Whites, those for non-owners to be higher than those for owners, and those for low return areas to be higher than those for high return areas. Although there are some exceptions for individual post-stratum comparisons, the first two expectations are very

well satisfied as shown in Table 2. The expectation that low return rate post-strata should have higher CCFs than the corresponding high return rate post-strata is met but not nearly as convincingly: 1.0197 over all low return areas and 1.0167 over all high return areas. This difference is not statistically significant and it should be noted that not all PES persons could be coded. Although return rate is clearly not as important as the other post-stratification variables, the low return areas do have 40% more problem records (erroneous enumerations plus nonmatches) than the high return areas, indicating the potential value of the return variable. Changes in collection procedures and type of enumeration area between 1990 and 2000 and the larger A.C.E. sample sizes may lead to more substantial results for the return variable in 2000.

Table 2: Synthetic Coverage Correction Factors for Selected Subpopulations Defined by the Post-stratification Variables

Lowest CCF		Highest CCF	
Owners	1.0008	Non-Owners	1.0513
N-H Whites	1.0076	Others	1.0491
50+ Female	0.9889	18-29 Male	1.0391
High Return	1.0167	Low Return	1.0197
Non-Hispanic White or Some other race			
Small MO/MB	1.0025	Update/Leave	1.0130
NE owners	0.9898	South owners	1.0047
Blacks and Hispanics			
Small or U/L	1.0455	Large/Medium	1.0541

Graph 1 shows the synthetic CCFs for the PES and A.C.E. designs and 95% confidence intervals (based on the PES data) for the two designs for the 51 PES post-stratum groups and selected subtotals defined by race/Hispanic origin and tenure. The standard errors of the difference of the two estimates is usually approximately the larger of the two standard errors. The A.C.E. confidence intervals are narrower mainly because of the borrowing of strength from many A.C.E. post-strata for the estimates for a single PES post-stratum. The A.C.E. estimates for most of the PES post-stratum groups are within the PES confidence intervals, indicating that the proposed A.C.E. post-stratification design is consistent with the 1990 PES design and that the A.C.E. post-stratification is not adversely affecting the heterogeneity bias. For the discrepancies, the estimated PES and A.C.E. coverage correction factors for Hispanic owners in large urbanized areas in the Midwest, a very small cell, are beyond sampling error, but the PES estimate, about 0.96 with SE 0.024, is 0.06 or more lower than for other Hispanic owners. The source of the discrepancy is almost certainly in the small size of the PES post-strata. The absence of region in the A.C.E. design for minorities prevents this anomaly. Similarly, the estimated PES coverage correction factors of about 1.06 (SE 0.021) for Black owners in large urbanized areas of the West and 1.19 (SE 0.073) for Hispanic non-owners in non-urban areas seem out of line compared to similar post-strata. The standard errors that should be expected for the A.C.E. for Census 2000 should be smaller than those shown here because of the increased A.C.E. sample size and improved weighting.

Graph 2 shows the same data as Graph 1 for the states and the District of Columbia. The standard errors of most states are reduced, some substantially. The state estimates for the proposed A.C.E. post-stratification design are generally within sampling error of the PES estimates except in the Midwest where the large number of non-urban households generally have good coverage in this region but poorer coverage elsewhere. The absence of the regional post-stratification for Non-Hispanic White non-owners, Blacks, and Hispanics results in the application of generally higher national coverage factors. The higher estimated coverage correction factors in the Midwest are offset by lower coverage correction factors in the Mountain states. The Mountain states' estimates with the proposed A.C.E. post-stratification design are still within sampling error of the PES estimates.

In comparing the PES and A.C.E. designs, no congressional district's population changes by more than the larger of 10,000 or 1.75 percent, and no congressional district's share of a thousand dollars changes by more than the larger of 4 cents or 1.75 percent.

There are more post-strata for Non-Hispanic White owners in the A.C.E. design than in the PES design. The standard errors for A.C.E. post-strata with up to 1,500 P-Sample persons in the PES have lower average standard errors than similar size PES post-strata. Above 1,500 P-Sample persons, the situation is reversed. Because of collapsing, there are fewer post-strata for the other population subgroups. Controlling for race, similar size post-strata have similar A.C.E. or PES average standard errors.

Several sets of target values were simulated in order to estimate bias and mean square error. These did not assist in discriminating between post-stratification designs and were dropped.

Conclusion

The proposed census 2000 A.C.E. post-stratification design is an appropriate design, based on the data from the 1990 PES, for measuring coverage of Census 2000.

- Larger sample sizes, decreased weight variability and more aggressive collapsing of small post-strata mean

that estimated A.C.E. standard errors will be lower than those from the 1990 PES design.

- Two additional variables, tract level return rate and type of enumeration area, will be incorporated where possible to help control heterogeneity bias.
- The estimated A.C.E. coverage correction factors are generally within sampling error of the 1990 PES design coverage correction factors.

References

Haines, Dawn E. and Hill, Joan M. (1998) "A Method for Evaluating Alternative Raking Control Variables." *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, pp. 647-652.

Hogan, Howard (1993) "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 1047-1060.

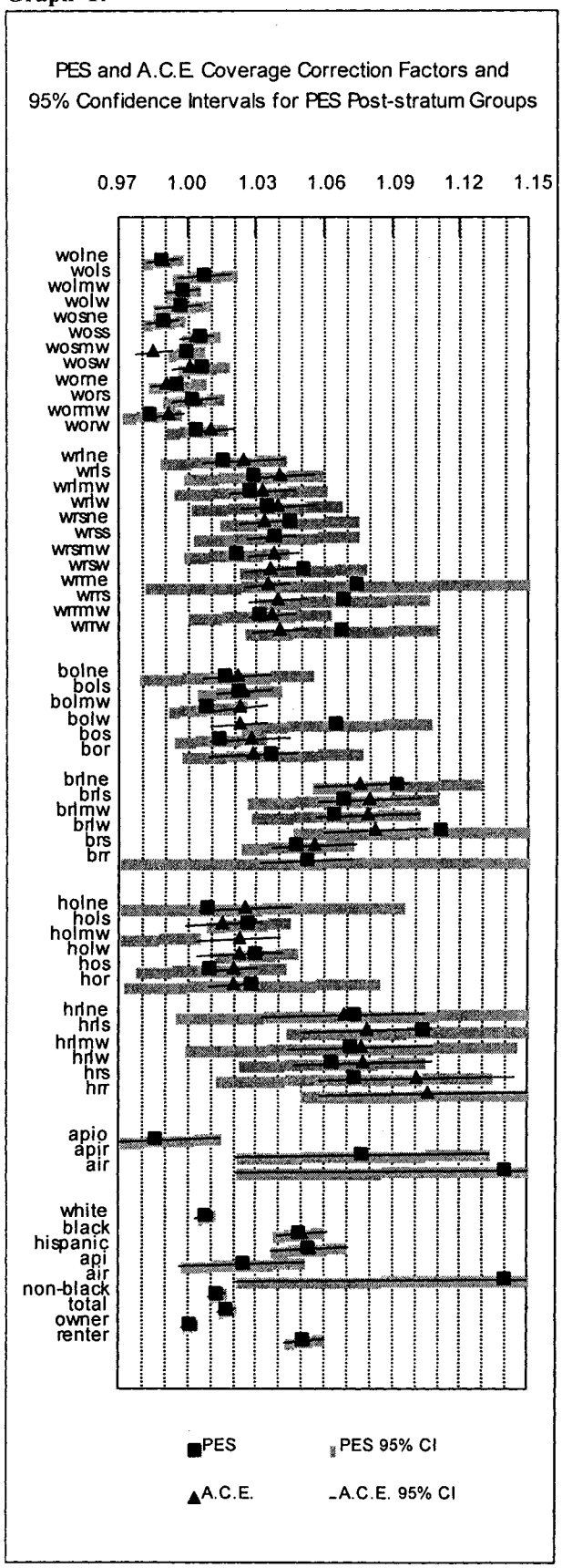
Schindler, Eric (1999) "Iterative Proportional Fitting in the Census 2000 Dress Rehearsal." *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, pp. 450-455.

Appendix A: Major Post-stratum Groups for 448 Post-strata

Persons are generally assigned to the lowest numbered applicable race/Hispanic origin domain. Each post-stratum group will have 7 age/sex post-strata for a total of 448 post-strata. Post-strata with less than 100 persons in the P sample will be collapsed by age and sex.

Race / Hispanic Origin	Tenure	MSA/TEA	High Return Rate				Low Return Rate			
			NE	MW	S	W	NE	MW	S	W
Non-Hispanic White or "Some other race" (7)	Owner	Large MSA MO/MB	1	2	3	4	5	6	7	8
		Medium MSA MO/MB	9	10	11	12	13	14	15	16
		Small MSA & Non-MSA MO/MB	17	18	19	20	21	22	23	24
		All Non- MO/MB	25	26	27	28	29	30	31	32
	Non-owner	Large MSA MO/MB	33				34			
		Medium MSA MO/MB	35				36			
		Small MSA & Non-MSA MO/MB	37				38			
		All Non- MO/MB	39				40			
Non-Hispanic Black (4)	Owner	Large & Medium MSA MO/MB	41				42			
		Small & non-MSA & all Non-MO/MB	43				44			
	Non-owner	Large & Medium MSA MO/MB	45				46			
		Small & non MSA & All Non- MO/MB	47				48			
Hispanic (3)	Owner	Large & Medium MSA MO/MB	49				50			
		Small & non-MSA & All Non- MO/MB	51				52			
	Non-owner	Large & Medium MSA MO/MB	53				54			
		Small & non-MSA & All Non- MO/MB	55				56			
NH/PI (5)	Owner		57							
	Non-owner		58							
Non-Hispanic Asian (6)	Owner		59							
	Non-owner		60							
AI/AN	Reservation (1)	Owner	61							
		Non-Owner	62							
	Not Reservation (2)	Owner	63							
		Non-owner	64							

Graph 1:



Graph 2: