A METHOD FOR EVALUATING ALTERNATIVE RAKING CONTROL VARIABLES

Dawn E. Haines, Joan M. Hill, Bureau of the Census
Dawn E. Haines, Bureau of the Census, Washington, DC 20233

KEY WORDS: Coverage Factor, Dual System Estimation, Iterative Proportional Fitting

ABSTRACT: Population coverage error estimates for the 1990 Decennial Census were based on Dual System Estimation (DSE) where one system was the census enumeration and the second system was an enumeration for a sample of the population as part of the Post Enumeration Survey (PES). Population coverage error estimates were based on 357 poststrata. Results from PES poststrata estimation indicated that differential undercounts existed across race and ethnic groups, renters, and rural residents. Iterative proportional fitting, or raking, will be used for the Census 2000 Dress Rehearsal to produce acceptable site-level estimates. The raking method corrects initial phase estimates by controlling to dual system estimates. Earlier research shows that increasing the number of poststrata and allowing multiple dimensions in the raking matrix yields more accurate coverage probabilities than DSE without raking. Our research focuses on constructing the best raking matrix for obtaining an accurate population estimate. We use logistic regression models to determine the optimal marginal, or control, variables. We then decide the dimensions and the placement of the variables on the raking matrix. Finally, we compare the performance of alternative raking matrices using coverage factor coefficients of variation and mean square errors.

I. Introduction

The 1990 PES collected information from randomly selected block clusters across the United States. The census enumerations in these block clusters comprise the E sample. The P sample is an independent listing of the chosen block clusters. The E sample is used to estimate erroneous inclusions in the census while the P sample is used to estimate the number of persons not captured in the census. Using these two samples and the census, direct and synthetic estimates of the population coverage error were constructed under DSE assumptions based on 357 poststrata (Mulry et al. 1997). Poststrata were defined by region, race, tenure, age/sex, and urbanization.

Currently, there are two planned phases for Census 2000. The initial phase enumerates persons via mail and personal visits while the second phase consists of the Integrated Coverage Measurement (ICM) quality check survey. The ICM survey, which presently consists of approximately 750,000 housing units, corrects the initial phase census estimate for coverage errors. An ICM sample is selected from each individual state in order to satisfy the requirement that state population estimates are based only on that state's data. According to Schindler and Griffin (1997), acceptable state estimates can be produced for approximately thirty poststrata per state under current sample size restrictions using DSE without raking.

In order to control on more variables without increasing the variance of the estimates (due to additional poststrata), a new poststratification was proposed for producing census population estimates. The proposed poststratification uses iterative proportional fitting, or raking, to account for persons not captured in the initial phase of the census. Raking has been used in past censuses to weight long form data to produce sample item estimates. Raking is typically used in survey work to adjust the results of a survey to match decennial census marginal counts which include adjustments for births, deaths, immigration, and emigration. However, if raking is used in Census 2000, it would correct initial phase estimates to match dual system estimates. Schindler and Griffin (1997) show that allowing a greater number of poststrata and/or multiple dimensions in the raking matrix yields more accurate coverage probabilities than the traditional DSE.

II. General Methodology

The objective of this paper is to compare the results of raking using proposed marginal variables (marginals) to raking results using variables derived from logistic regression models. Logistic regression modeling is a mechanism for determining variables which explain inclusion in the initial phase of the census. We also compare raking results to poststratification without raking. Different sets of marginals are evaluated based

---

The authors are mathematical statisticians in the Decennial Statistical Studies Division and the Planning, Research, and Evaluation Division, respectively. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

on their coverage factor coefficients of variation (CVs) and mean square errors. A coverage factor, which is the ratio of the DSE to the initial phase estimate at the poststratum level, is produced for each interior cell of the raking matrix. Multiplying the coverage factor by the initial phase estimate at the block level yields a synthetic estimate which corrects for coverage.

Raking methodology entails calculating a DSE for each interior cell. Summing the DSEs over rows and columns yields marginal controls. The initial phase estimates are then raked to the marginal controls. Direct DSEs are not calculated for the marginals because it leads to high correlation bias. For example, a DSE for black males age 18-29 groups both owners and renters together.

State and substate estimates are produced using California data from the 1990 PES. American Indians living on reservations are excluded from our study because sampling and estimation procedures for the 1990 and 2000 censuses differ for this population (Schindler and Griffin 1997). California was selected for this study because its 1990 PES sample size is approximately equal to the ICM sample size for a typical small state in Census 2000.

The approach used to assess alternative raking control variables is summarized as follows. First, potential raking variables are identified. Alternative logistic regression models are developed using the proposed raking variables. Once significant model variables are selected, they are combined to form a raking matrix. The same set of variables could lead to raking matrices with different sizes and dimensions. Next, raking is implemented which adjusts the 1990 census counts to match the 1990 DSEs based on the given marginals. Coverage factors are calculated for each alternative model at the state and substate level. The mean and range of the coverage factor CVs are compared to determine the best marginal variables. The jackknife procedure is used to produce standard errors of the coverage factors. One block cluster is removed at a time and the full raking process is repeated.

III. Description of Methodology

A. Variable Definitions

An assortment of variables are considered as inputs to the logistic regression model. The explanatory variables used to rake population estimates are those most highly correlated with whether or not a person is enumerated in the initial phase of the census. Known correlates such as race, hispanic origin, age, sex, and tenure are proposed as well as variables not previously considered. Specifically, we consider the following variables: race/hispanic origin, age/sex, tenure, family stability, urbanicity, family composition, percent non-owner, mail response rate, percent minority, vacancy rate, household size, and relationship. The intercept is always included in the logistic regression models. Following is a description of each independent variable.

Race/Hispanic Origin: (1) non-Hispanic white or other, (2) black, (3) Hispanic white or other, (4) American Indians not on reservations, and (5) Asians and Pacific Islanders.

Age/Sex: (1) under 18, (2) male 18-29, (3) female 18-29, (4) male 30-49, (5) female 30-49, (6) male 50+, and (7) female 50+.

Tenure: (1) owner and (2) renter.

Family Stability: (1) stable and (2) not stable. A household is "stable" if either (1) there is only one resident and that resident is over age 50, or (2) there are two to seven residents, the first two residents are over age 30 and are of the opposite sex, and any additional residents are under age 18. Households with more than seven residents or with any resident between 18 and 29 are never deemed "stable."

Urbanicity: (1) non-urban area and (2) urban area.

Family Composition: (1) spousal and (2) non-spousal. A household is considered "spousal" if one of two conditions are met: (1) the second person listed on the census form is the spouse of the householder or (2) the unit is occupied by exactly one person and that one person is over the age of 50.

Percent Non-Owner: (1) high and (2) other. Percent non-owner is a block-level variable. High percent non-owner blocks are those with greater than 65.73 % (75th percentile) non-owners.

Mail Response Rate: (1) low and (2) other. A block-level variable defined as the proportion of households in the 1990 mail universe which completed their 1990 Census form without the aid of an enumerator. Low mail response rate blocks are those with a mail response rate less than 53.25 % (25th percentile).

Percent Minority: (1) high and (2) other. Percent minority is a block-level variable. High percent minority blocks are those with greater than 79.17 % (75th percentile) minorities.

Vacancy Rate: (1) high and (2) other. Vacancy rate is a block-level variable. A block vacancy rate is high if it is greater than 6 %. This rate is not based on quartiles because of the small range of vacancy rates. A natural cutoff rate of 6 % occurred in the data.

Household Size: (1) one and (2) two or more.

Relationship: (1) related to person 1 and (2) not related to person 1. This distinction is made for all persons listed on the census form.

The dependent variable in the logistic regression model is a person-level indicator for inclusion in the E (or P) sample. For the E sample model, each E sample person is classified into one of three categories. If an E sample person matches a person in the P sample, the case is called a match and the dependent variable is assigned the value 1. If a person is found in the P sample but does not match any record in the E sample, the dependent variable is assigned the value 0. This case is called a P sample nonmatch. Finally, an E sample nonmatch case occurs when a person is in the E sample but does not match any record in the P sample. The value 1 is assigned to the dependent variable in this case. Similar definitions exist for the P sample model. We can view the results of the E and P sample models as two independent assessments of the population.

B. Model Development

Much work has been done over the years to identify characteristics of persons not enumerated in the census. Since the majority of the nation is enumerated using a mail census, a large body of literature exists on nonresponse to the census mail questionnaire. The E sample model focuses on correlates of being captured in the census, or E sample, since the dichotomous capture variable is the dependent variable in the logistic regression model.

Alho et al. (1993) develop logistic regression models of capture probabilities which allow person-, household-, and block- level characteristics as explanatory variables in addition to geographic factors such as urbanization. The logistic approach permits the use of continuous explanatory variables. The E and P samples are used to identify those persons captured only in the E sample, only in the P sample, and in both samples. Categorizing persons in this manner was a major challenge in the 1990 PES since it was not designed to provide this type of information at the person level (Alho et al. 1993).

The logistic regression approach to obtaining correlates of capture is complicated by the presence of unresolved cases since their match status could not be determined. For this reason, Alho et al. (1993) exclude unresolved cases and develop models based only on resolved cases. This does not introduce bias provided the probability of being captured is not significantly different between resolved and unresolved cases for a given model.

Mulry, Davis, and Hill (1997) use similar logistic regression models to estimate the capture probabilities of persons in the 1990 Census. Their paper studies the feasibility of using estimated probabilities to model heterogeneity in census coverage error for small areas. They examine several main effects and two-factor interactions not considered by Alho et al. (1993).

To determine the most appropriate estimation design, we compare poststratification estimates to a number of raked estimates using various potential raking variables. Our study and the Census 2000 Dress Rehearsal model define poststrata by race, hispanic origin, age, sex, and tenure. Our study utilizes a 35x2 raking matrix which cross-classifies race/hispanic origin and age/sex on one margin while tenure is on the other margin. The raking matrix for the Census 2000 Dress Rehearsal is 42x2 because the race variable has been expanded to six categories. Historically, these variables are highly significant and will most likely remain in the final raking matrix for Census 2000.

With this in mind, alternative raking models are formed by adding one additional variable to those already deemed significant. In this case, logistic regression modeling identifies another variable to include in the model already containing race/hispanic origin, age/sex, and tenure. The odds ratio and a Wald chi-square test statistic are used to determine which explanatory variable to include in the model. The odds ratio for a given explanatory variable is computed by exponentiating the parameter estimate. The Wald chi-square test statistic is computed as the square of the value obtained by dividing the model parameter estimate by its standard error. The significance level used in this study is $\alpha = 0.10$ or 90 percent significance. Model results are obtained using SAS PROC LOGISTIC which assumes the data are produced using simple random sampling. As a result, the chi-square critical value for $\alpha = 0.10$ and 1 degree of freedom is multiplied by a design effect (DEFF) of 20.2 which accounts for the complex sample design of the PES when estimating standard errors. The DEFF used for this study is the ratio of the variance of the 1990 Census population undercount percentage under the PES design and the variance of that statistic assuming simple random sampling. The variance of the undercount based on the complex PES sample design and jackknife replication variance estimation methodology is the square of the

standard error (i.e., the square of 0.45%). The variance for the undercount rate estimate of 0.04 assuming simple random sampling is approximately 1.0 x 10E-6. Thus, the ratio of the variances (DEFF) is approximately 20.2 while the ratio of the standard errors (DEFT) is approximately 4.5. The three best variables based on logistic regression modeling are relationship, urbanicity, and mail response rate.

Given alternative sets of raking marginals, we must decide how to place the variables on the raking matrix. That is, how will variables be grouped on each of two or more raking margins? Ideally, marginal variables which are correlated will be cross-classified within the same dimension. Three-(or more) dimensional raking is possible. Increasing the number of dimensions is expected to result in better direct estimates for the marginals, but collapsing to eliminate cells with little or no sample cases becomes much more difficult (Schindler and Griffin 1997). For each alternative model, it is possible to propose several configurations of the raking matrix.

C. Model Comparisons

Coverage factor estimates and their standard errors are computed using jackknife replication variance estimation methodology. One of the 383 California PES block clusters is dropped at a time to form each subsample. The estimate of interest is calculated from the full sample as well as from each subsample. The variation among the subsample estimates is then used to estimate the variance for the full sample. The form of the jackknife variance estimator is a constant times the sum of squared differences between the estimator based on the full sample and each subsample. We ignore the constant because it has value 382/383 = 0.9974.

IV. Results

Schindler and Griffin (1997) present notation for calculating raking coverage factors and their standard errors. Each poststratum is defined by the demographic subgroup represented by the corresponding cell of the raking matrix. Coverage factors are calculated for each interior cell of the raking matrix. Alternative raking models are evaluated by computing the CV of each interior cell coverage factor and then averaging the coverage factor CVs over all nonempty interior cells.

Table 1 compares the average, minimum, and maximum coverage factor CVs for both poststratified and raked estimates for six models. The first model is the proposed 35x2 model while the other five models are

Table 1: Coverage Factor CV Estimates for Poststratification and Raking for Alternative Models

Matrix (number of nonzero cells in parentheses)	Poststratified Estimates			Raked Estimates		
	Average	Min	Max	Average	Min	Max
Race/Orig X Age/Sex by Tenure (58)	0.040	0.006	0.106	0.030	0.006	0.065
Race/Orig X Age/Sex by Tenure X Urban (108)	0.141	0.000	1.332	0.041	0.006	0.079
Race/Orig X Age/Sex by Tenure X % MRR (116)	0.073	0.006	0.659	0.034	0.006	0.066
Race/Orig X Age/Sex by Tenure X Rel (116)	0.162	0.000	1.388	0.059	0.006	0.212
Race/Orig X Tenure by Age X Tenure X Rel (79)	0.111	0.000	0.999	0.054	0.004	0.134
Race/Orig X Tenure by Sex X Tenure X Rel (40)	0.073	0.000	0.258	0.044	0.005	0.079

alternatives based on logistic regression. The proposed 35x2 raking model has the smallest average coverage factor CV for both poststratified and raked estimates, though the other model averages are not substantially higher for raked estimates. In all cases, the average coverage factor CV for the raked estimates are lower than those for the poststratified estimates. In order to compare each model's overall effectiveness in reducing the mean square error of the population estimates, we consider each model's contribution to bias. The primary advantage of raking is to reduce the variance of the population estimates without substantially increasing the bias. Thus, we may be willing to accept a model with biased estimates if it yields a reduction in variance.

Table 2 displays the square root of the average coverage factor mean square error for poststratified and

raked estimates for the six models. For variance and bias estimates, we assume that the direct poststratified DSE is unbiased at the poststratum level. An estimate of the MSE of the raked coverage factor for poststratum p is given by $(CF_{p,rake} - CF_{p,direct})^2 - SE^2(CF_{p,rake} - CF_{p,direct}) + SE^2(CF_{p,rake})$. The square root of the weighted and unweighted poststratum average is presented in Table 2. Our study uses the weight $(CEN_p^2 / \sum CEN_p^2)$ where CEN is the census count and the summation is over all p poststrata. The weighted estimator prevents poststrata with small sample sizes from having a disproportionate effect. Assuming an unbiased DSE at the poststratum level ignores model bias, correlation bias, and the ratio estimation bias inherent in dual system estimation. As a result, we are measuring bias caused primarily by the raking procedure. The asterisks (*) in the raked bias columns indicate a negative estimate of squared bias. Bias estimates are not defined for these cases. Although bias sometimes exists in the raked estimates, the raked MSEs are always lower than the poststratified MSEs. This is true for both weighted and unweighted estimates.

Table 2: Square Root of Average Coverage Factor MSE for Poststratification and Raking for Alternative Models

Raking Matrix (number of nonzero cells in parentheses)	Weighted			Unweighted		
	Post-stratified MSE	Raked MSE	Raked Bias	Post-stratified MSE	Raked MSE	Raked Bias
Race/Orig X Age/Sex by Tenure (58)	0.0232	0.0189	0.0077	0.0506	0.0371	0.0086
Race/Orig X Age/Sex by Tenure X Urban (108)	0.0257	0.0194	0.0049	2.3197	1.6227	1.6219
Race/Orig X Age/Sex by Tenure X % MRR (116)	0.0262	0.0148	*	0.1308	0.0078	*
Race/Orig X Age/Sex by Tenure X Rel (116)	0.0282	0.0175	*	0.6073	0.3032	0.2917
Race/Orig X Tenure by Age X Tenure X Rel (79)	0.0195	0.0134	*	0.4298	0.2552	0.2419
Race/Orig X Tenure by Sex X Tenure X Rel (40)	0.0150	0.0128	*	0.1377	0.0595	*

Table 3 presents estimates for the total population coverage factor CVs over 383 block clusters in the California PES. Poststratified and raked estimates are compared based on the average, minimum, and maximum CV of the total population coverage factor. For each model, we compute coverage factors for all nonempty interior cells. Synthetic estimates and coverage factors are then calculated for each block cluster. Jackknife standard errors are calculated for the coverage factors by removing one block cluster at a time. Finally, the coverage factor CVs are averaged over all 383 PES block clusters in California.

Table 3 shows that the proposed 35x2 model with raking yields the smallest average total population coverage factor CV, although the corresponding estimates for the other models are not much higher. The mean of the raked estimates is less than or equal to that of the poststratified estimates for each model.

For each raking matrix, direct synthetic and raked estimates are compared to a target estimate. The target used in our example is constructed by combining all E and P sample people but removing the matches and erroneous enumerations. The average relative root mean square errors (RRMSEs) for the poststratified and raked estimates are given in Table 4 where the average is taken over all 383 block clusters. In general, the average RRMSE for the raked estimates are lower than those for the poststratified estimates, although they are very similar. This relationship does not hold for the last model where the raking average RRMSE is slightly higher.

Table 3: Total Population Coverage Factor CVs for Poststratification and Raking Over 383 Block Clusters

Raking Matrix (state level coverage factor CV in parentheses)	Poststratified Estimates			Raked Estimates		
	Average	Min	Max	Average	Min	Max
Race/Orig X Age/Sex BY Tenure (0.0042)	0.010	0.000	0.039	0.009	0.000	0.035
Race/Orig X Age/Sex BY Ten X Urban (0.0043)	0.015	0.000	0.330	0.011	0.000	0.049
Race/Orig X Age/Sex BY Ten X % MRR (0.0045)	0.015	0.000	0.069	0.011	0.000	0.039
Race/Orig X Age/Sex BY Ten X Rel (0.0049)	0.014	0.000	0.070	0.010	0.000	0.050
Race/Orig X Tenure BY Age X Ten X Rel (0.0044)	0.011	0.000	0.060	0.010	0.000	0.044
Race/Orig X Tenure BY Sex X Ten X Rel (0.0045)	0.010	0.000	0.055	0.010	0.000	0.058

Table 4: Average Relative Root Mean Square Error for Poststratification and Raking Over 383 Block Clusters

Raking Matrix	Poststratified Estimates	Raked Estimates
Race/Orig X Age/Sex BY Tenure	0.0565	0.0563
Race/Orig X Age/Sex BY Ten X Urban	0.0584	0.0567
Race/Orig X Age/Sex BY Ten X % MRR	0.0577	0.0567
Race/Orig X Age/Sex BY Ten X Rel	0.0604	0.0596
Race/Orig X Tenure BY Age X Ten X Rel	0.0595	0.0592
Race/Orig X Tenure BY Sex X Ten X Rel	0.0600	0.0603

The raking procedure allows consideration of a large number of poststrata without increasing the variance of the estimates. Although raking is useful for decreasing the variance of our estimates without substantially increasing the bias, raking yields only slightly better results than poststratification. We conclude that the Census 2000 Dress Rehearsal model is the best model based on the 1990 California PES data. Our goal is to implement this modeling procedure for all states in order to determine the best raking models for Census 2000.

V. Acknowledgments

The authors thank Eric Schindler for calculating estimates shown in the tables. We also thank Richard Griffin and Robert Fay for their helpful comments.

VI. References

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

Mulry, M. H., Davis, M. C., and Hill, J. M. (1997). A Study in Heterogeneity of Census Coverage Error for Small Areas. *American Statistical Association Proceedings of the Survey Research Methods Section.*

Schindler, E. and Griffin, R. (1997). Census 2000 ICM: Stratification and Poststratification. *American Statistical Association Proceedings of the Survey Research Methods Section.*

Word, D. (1997). Who Responds/Who Doesn't? Analyzing Variation in Mail Response Rates During the 1990 Census. *Working Paper No. 19 in the Working Paper Series of the Population Division, U. S. Bureau of the Census.*