SOME SMALL AREA ADJUSTMENT METHODOLOGIES APPLIED TO THE 1980 CENSUS

Gregg J. Diffendal, Cary T. Isaki, and Donald Malec, Bureau of the Census

The Census Bureau has instituted several programs for measuring the quality of the 1980 census especially the undercount of the population. Demographic analysis (DA) and the Post Enumeration Program (PEP) are the two major programs to estimate the 1980 undercount. However, DA provided population estimates of the legal population at the national level (states may be available later) while PEP, a sample survey, was designed to provide population estimates at states and some major SMSA's. Using data from DA and the PEP, several methods for adjusting 1980 census county total population are illustrated.

The measurements of the undercount from DA, PEP, and differential undercount measurements from previous censuses (mostly from DA) have created interest, both politically and statistically, to adjust the 1980 census of its measured errors. Politically, the many lawsuits filed for adjustment of the 1980 census against the Census Bureau are ample evidence. Two of the many lawsuits, by Detroit and Philadelphia, are described by Barabba et. al. (1983). Congress even passed a law stipulating that the Census Bureau could use adjusted census figures in the calculations used for the general revenue-sharing estimates. Statistically, an undercount conference was held where many papers were presented and models were proposed for adjusting the census.

As mentioned, the two ways of measuring the undercount are DA and the PEP. DA has produced undercount estimates by age-race-sex since 1950. DA is essentially an accounting technique. It estimates the population for 1980 by taking a previous census (adjusted for its undercount) and adding births, subtracting deaths, adding immigrants, and subtracting emigrants. For the older age groups these estimates are supplemented by social security records which are believed to be nearly complete. In equation form, DA estimates are

$$P_1 = P_0 + B - D + I - E$$

where P is for population, B births, D deaths, I immigrants and E emigrants. Birth and death records are fairly complete, but the estimates of immigrants and emigrants are questionable. Legal immigrants estimates can be obtained from the Immigration and Naturalization office, but there are an expected large and unknown number of illegal aliens that may not have been counted in the census. The number of emigrants is believed to be small but is also difficult to estimate. The data used in the illustrations that follow from DA is broken down by age, race (Black, Non-Black) and sex at the national level using the estimates from U.S. Bureau of the Census (1982). Approximately 2.5 million illegal aliens were added to these estimates so that the estimate of the population at the national level would equal the PEP estimates. Two different assumptions of the race of the illegal aliens were used which will be discussed later.

The PEP estimates are obtained from a sample survey that is matched to the census. Two samples were taken, one in April and one in August, both were essentially CPS samples. Only the sample estimates known as 1-7 are used here. These are the preliminary estimates derived from the April sample before clean-up operations. The PEP uses a dual system procedure in order to estimate the population total. This procedure requires an assumption of independence between the census and the PEP, an assumption that may not hold in practice. Missing data in the PEP has led to the development of several methods for imputation and hence differing estimates of the undercount. See Cowan and Bettin (1982) for a full discussion of the PEP and the assumptions made.

Unlike DA, the PEP does provide undercount estimates of the non-institutional population by states, large SMSAs, and some major cities by age-race (White, Black, Other by Hispanic/Non-Hispanic)-sex. Only results using the state estimates are presented here. Also, the PEP does not require a separate estimate of the illegal population, but it does estimate the institutional population separately with less detail than the estimates of the noninstitutional population.

The count for the 1980 census was 226.5 million people. PEP estimates the total population to be 228.1 million people. Therefore the national undercount estimate used here estimates the undercount at 1.6 million people. Note that the variable reported in the discussion that follows is the population estimate divided by the 1980 census counts so the national undercount ratio for all methods described is 1.007.

The methods of adjustment used below are the synthetic and regression methods. The assumptions of these methods will be made clear. First the DA and PEP populations estimates are assumed given at the national and state levels, respectively. Therefore, the reliability of the DA national estimates and PEP state estimates should be discussed separately of this work, although the quality of the small area estimate can be no better than the DA or PEP estimates.

The synthetic estimator, essentially a ratio estimator, assumes the national (for DA) or state (for the PEP) undercount rates by age-race-sex hold at the smaller geographical levels by age-race-sex. Mathematically, the synthetic estimator for DA can be written as

$$P_m = \sum_{(i,j,k)} C_m(ijk)[D(i,j,k)/C(i,j,k)]$$

where

$D(i,j,k)$ denotes the DA data for age group i, race j, and sex k at the U.S. level.

$C_m(i,j,k)$ denotes the comparable census total for the m^{th} county.

$C(i,j,k) = \sum_m C_m(i,j,k)$

where $\sum_m$ denotes summation over all counties in the U.S. When the PEP data was used D(i,j,k) is defined for each state separately and applied only to the counties within a state.

The regression estimator assumes that a linear combination of independent variables explains the undercount for the small areas. The regression coefficients that are calculated on state data are assumed to hold for the small areas (counties). This is analogous to the assumption that national undercount rates, for a specific age-race-sex cell, apply at smaller areas for the synthetic estimator. Since the PEP is a national sample, theoretically we can obtain sample estimates at smaller levels of aggregation, such as PSU, counties or others for a sample of such levels. In this manner the model will be fitted on units more comparable to units we wish to estimate. This approach will be tried at a later time, although there are technical difficulties, such as biased estimates, larger variances and handling of missing data required for the dual system estimate used to estimate the undercount. Some of these difficulties may prove to be insurmountable. Three different regression estimators will be discussed: ordinary least squares, weighted least squares, and a maximum likelihood estimator that accounts for model and sampling variances.

Six adjustment methodologies will be discussed: two using DA and four using the PEP. Since the two DA estimates and the PEP result in different state estimates the state estimates will be discussed first. All four PEP county estimates have the same state totals. The methods are summarized in the Table.

Method A1 is a synthetic estimator which uses the DA data. Race is defined as Black-Non-Black only. The higher undercount for Blacks as measured at the national level are translated into higher undercounts for the states in the South. Only the states in the South have undercount ratios above 1.01, all other states are between .99 and 1.01 showing little change from the census. If we were to adjust the 1980 census using Method A1 then the change in apportionment from using the census data for the House of Representatives would be that Georgia would gain one seat and New York would lose one seat.

Method A1 assigns the illegal aliens to the Non-Black race category. This is a major flaw of method A1 because the majority of the illegal aliens are believed to be Hispanic. Method A1-MOD2 assigns 95% of the 2.5 million illegal aliens to the Hispanic race category and the other 5% to the Black race category (the same as Method A1). Therefore, there are now three race categories: Black, Hispanic, and Other. Some assumptions were made to divide the Non-Black category into the Hispanic and Other categories. The results show that the lower South, the Western states along the Mexican border, and New York have undercounts of 1.01 or higher with Texas and New Mexico over 1.03. The Dakotas and most of New England have overcounts below .99. If Method A1.MOD2 was used to adjust the 1980 census, then California and Texas gain one seat each and Ohio and Pennsylvania would lose one seat each.

The PEP state estimates used in the illustrations are the preliminary estimates based on households in the April CPS sample. The results from the PEP show an undercount in almost every state in the West, the middle of the country has little change from the census, and the South and Northeast areas are mixed. The two extreme states with their undercount ratios in parenthesis are South Carolina (1.078) and Tennessee (.97). If the PEP estimates were used for apportionment in the House of Representatives, then California would gain one seat and Pennsylvania would lose one seat.

The six different county estimates will now be discussed. The Table below summarizes the six different estimators. All county estimates sum to their state totals. That is, the estimators of county population utilizing data from the PEP are ratio adjusted to the PEP state total. Since DA and PEP differ in their state estimates, some differences are due to the state totals and not to the methodology used.

DA synthetic method A1 results for the counties are similar to the results discussed previously for the states. The counties in the South have high undercounts which coincide with the counties with high percent Black. There is a strong linear relationship of percent Black with the undercount ratio. The rest of the country has undercount ratios around 1.00 showing very little difference from the census counts.

DA synthetic method A1-MOD2, which uses Hispanic, Black and Other as the race category, results show that areas with high Hispanic and high Black populations have high undercount ratios. The counties along the Mexican border, Texas, New Mexico, Arizona, and California join the counties in the South with high undercount ratios. Some major cities outside of these areas also have high undercount ratios, such as Chicago and New York. Many counties along the Canadian border and generally the Northeast and central plains have sizable overcounts, with undercount ratios below .99.

PEP synthetic method B1 shows results similar to the PEP state estimates but with more differences within a state than the DA methods. Most counties within Vermont, Connecticut, New York, Pennsylvania, Kentucky, Tennessee, and Alabama have sizable overcounts with undercount ratios below .99. Counties in South Carolina, Wyoming, Nevada, and California have high undercount ratios. Other counties in the West and Maine have fairly high undercount ratios (1.01 to 1.025) except for counties in Oregon, Utah, and Colorado. The counties along the Mexican border do show high undercount ratios, even in Texas, which coincide with method A1-MOD2. Counties in Texas not along the border of Mexico have undercount ratios below 1.01 and some show overcounts. This reflects the disproportionate numbers of Hispanics in counties along the Mexican border. Most other counties in the nation have undercount ratios around 1.00 (or close to

census counts) although some are higher or lower.

The PEP regression estimates show a striking difference from the synthetic estimates. Considerably more variability exists. PEP regression Method C1 is a linear regression with independent variables chosen by a stepwise procedure from a set of 29 variables such as race, substitution, allocations, housing, birth, death, etc. There are some problems with these variables for some counties-recording error and variables that are considerably higher or lower than any state variable. Most variables are expressed in percent form so that the range for the states is similar to those for the counties. All variables are expressed in a form which is independent of level. The estimated regression equation under model C1 is

$$\underset{(.13)}{Y = .807} + \underset{(.36)}{1.19 \cdot X1} - \underset{(1.8)}{5.17 \cdot X2} + \underset{(.13)}{2.16\ X3}$$

where Y is PEP undercount estimate/ 1980 census (mean = 1.006), X1 is percent substitution (mean = .015), X2 is 1979 deaths/ 1980 census (mean = .008) and X3 is 1979 Medicare/1980 census ages 66+ (mean = 1.04). The values in parenthesis below the regression coefficient are the standard errors. The estimated model variance is .00023. An R^2 = .26 was achieved with this model.

It is difficult to summarize the patterns of the undercount ratios for method C1. There is much more variability of the county ratios within a state. Since the counties sum to their state totals, states with very high undercount ratios (such as South Carolina = 1.07) and states with low undercount ratios (such as Tennessee = .97) generally have county ratios around these values.

PEP regression Method C2 used a weighted least squares estimator with the weights being the inverse of the sample variance. The estimated regression equation under model C2 is

$$\underset{(.017)}{Y = 1.05} \underset{(.164)}{-.487 \cdot X4} + \underset{(1.38)}{4.50 \cdot X5} \underset{(.17)}{-.257 \cdot X6} \underset{(1.37)}{-3.25 \cdot X2}$$

where X2 is 1979 deaths/1980 census, X4 is percent Hispanic (mean = .042), X5 is percent Hispanic allocations (mean = .006), X6 is percent white Non-Hispanic allocation (mean = .074). The variance is E_2W where W is the inverse of the sampling variance (mean = .00012) and E_2 is estimated to be 1.63.

For method C2 the undercount ratios are closer to method B1 although more variability does exist. The major difference is that the counties along the Mexican border have undercount ratios below .99 showing overcounts except in the counties in California which have high undercount ratios.

PEP regression Method C3 is a maximum likelihood estimator that assumed the model error is additive with the sampling error. An iterative procedure was used which assumes multivariate normality. Since no selection procedure is available all the variables from methods C1 and C2 were used. Note that Method C1 assumes no sampling error while method C2 assumes the model error is multiplicative with the sampling error. The estimated regression equation under model C3 is

$$Y = .816 + 5.94 \cdot X1 - 4.19 \cdot X2 + .219 \cdot X3$$

$$- .503 \cdot X4 + 4.60 \cdot X5 - .221 \cdot X6$$

where the terms are defined as before. No standard errors of the coefficients are available for this model. The error for model C3 is $E_3 + W$ where W is as before and E_3 is the model variance estimated to be .000019.

The undercount ratios for method C3 appear similar to those for method C2 although some differences do exist. Except in a few states (Maine, South Carolina, Ohio, and Wisconsin), many counties not in the West show undercount ratios below .99, overcounts.

In summarizing the county estimates, the synthetic estimates resemble fairly closely the results for the states. The regression estimates have much higher variability.

Further Work

The most questionable assumption made in the regression work is calculating the regression coefficients from state data and applying them to the counties. Work will be done to obtain PEP estimates at PSU, district office, and perhaps county level. We have already obtained PEP estimates for major central cities, large SMSA's, and balance of states. Other work contemplated or in various stages of completion are using robust regression; using prediction sum of squares and Mallows C_p as criteria of selection; obtaining other independent variables especially from the census long forms; using regression diagnostics to find influential data points and for variable selection; using principal components for reduction in the number of variables and removing near collinearity among the variables.

Can the 1980 census be adjusted? The work here is a first step toward answering this question. We do not think any of the models presented here are totally satisfactory. Even if a good model could be obtained, how would we know it is better than the census? Besides decisions on adjustment for the 1980 census which could be done for revenue sharing if not for the published official figures, this work may influence decisions on adjustment for the 1990 census. It may also influence the design of a PEP for census adjustment and how the census will be taken in 1990 if adjustment is an integral part of the enumeration process that produces the final published population counts.

Table. Description of County Adjustment Methods

COUNTY ADJUSTMENT METHODS	DATA SOURCE	ESTIMATOR	DESCRIPTION
A1	demographic analysis	synthetic	race-black, nonblack
A1-MOD2	"	"	race-black, hispanic, other
B1	PEP	"	state
C1	PEP	regression	OLS
C2	PEP	"	WLS
C3	PEP	"	MLE - model error and sample error

REFERENCES

Barabba, V. P., R. O. Mason and I. I. Mitroff (1983). Federal Statistics in a Complex Environment: The Case of the 1980 Census. The American Statistician, 37, No. 3, 203-212.

Conference on the Census Undercount, July, 1980, GPO Stock No. 80-607998.

Census of Population and Housing, 1980: Summary Tape file 2B (Internal File) [machine readable data file]/prepared by the Bureau of the Census,--Washington, D.C. 1982.

Census of Population and Housing, 1980: Summary Tape File 2 Technical Documentation/ prepared by the Data User Services Division, Bureau of the Census--Washington: The Bureau, 1982.

Cowan, C. D. and P. J. Bettin (1982). Estimates and missing data problems in the post enumeration program. Unpublished paper presented to the ASA technical panel on the undercount.

Diffendal, G. J., C. T. Isaki, and D. J. Malec (1982). Examples of some adjustment methodologies applied to the 1980 census. Unpublished paper presented to the ASA technical panel on the undercount.

Ericksen, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. Demography, vol. 10, no. 2, 137-160.

Ericksen, E. P. (1974). A regression method for estimating population changes of local areas. Journal of the American Statistical Association, 69, 867-875.

U.S. Bureau of the Census, Current Population Reports, series P-23, No. 115, "Coverage of the National Population in the 1980 Census by Age, Sex and Race: Preliminary Estimates by Demographic Analysis," U.S. Government Printing Office, Washington, D.C. 1982.

Warren, R. (1982). Estimation of the size of the illegal alien population in the United States. Unpublished paper presented at the annual meeting of the Population Association of America, San Diego, California.