ASPECTS OF SURVEY AND MODEL-BASED POSTCENSAL ESTIMATION OF INCOME AND POVERTY CHARACTERISTICS FOR STATES AND COUNTIES

Robert E. Fay and George F. Train

Presented at the
American Statistical Association Conference
Annual Meeting in Orlando, Florida on August 13-17, 1995

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

ASPECTS OF SURVEY AND MODEL-BASED POSTCENSAL ESTIMATION OF INCOME AND POVERTY CHARACTERISTICS FOR STATES AND COUNTIES

Robert E. Fay and George F. Train, U.S. Bureau of the Census[1]
Robert E. Fay, U.S. Bureau of the Census, Washington, DC 20233

KEY WORDS: *Empirical Bayes, small domain, small area, variance estimation*

Abstract: Congress has charged the Census Bureau with the responsibility to develop postcensal estimates of basic income and poverty statistics by state, county, and subcounty areas. The March Supplement to the Current Population Survey (CPS), based on a sample size of about 65,000 households, provides national estimates for these characteristics annually, but the sample size is insufficient to provide reliable direct estimates at all or even most levels of detail requested by Congress. The paper reports on two aspects of the estimation problem: 1) variance estimation for the CPS sample estimates, and 2) features of some candidate small area models that might be used to produce the required estimates at the state and county level. The two problems are interrelated: many small area models employ estimated variances for the survey estimates both in the estimation of the small area models and in their evaluation. The paper reports progress in developing a first set of estimates to meet the Congressional request.

1. Introduction

Although the Census Bureau has produced postcensal estimates of population for decades (Long 1993), Congress has often needed to rely upon statistics from the previous decennial census for characteristics such as poverty estimates and income distributions in block grants and other formula-driven distributions of funds to state and local governments. The reliance on past census data, which can become outdated, primarily reflects the absence of alternatives. The Census Bureau's program of Current Surveys generally has sample sizes adequate only for national rather than subnational analysis. For example, the Census Bureau was directed and funded to carry out a one-time 1976 Survey of Income and Education (SIE) in order to obtain state-level estimates of children in poverty. The SIE had a sample size approximately 3 times as large as the Current Population Survey (CPS), which has been the official source of annual national estimates of poverty and the income distribution.

For a period of time, the Census Bureau produced state and local estimates of per capita income, originally for the Revenue Sharing Program, using administrative data including information from tax returns. The Census Bureau has also produced annual estimates of median income for 4-person families by state based on a model combining CPS and administrative estimates (Fay, Nelson, and Litow 1993).

Congress has directed the Census Bureau to develop postcensal estimates of six characteristics at the state, county, and school district level. The characteristics are:

1) median household income,
2) per capita income,
3) children age 0-4 in poverty,
4) children age 5-17 in poverty,
5) persons age 65+ in poverty,
6) total persons in poverty.

The March Supplement of the CPS, which asks questions on income during the previous year, provides official national estimates annually.

This paper describes aspects of the research effort directed at state and county estimates of these characteristics. Section 2 describes our recent work to estimate variances and covariances for the CPS direct estimates at the state level. In general, obtaining accurate variance and covariance estimates is an important step in developing small domain estimates. Section 3 presents some initially encouraging findings for estimation of median income and per capita income at the state level.

2. Variance Estimation for the CPS

2.1 General Issues

Although the Census Bureau works within the design-based paradigm for analysis of survey data, direct calculation of variance estimates from the CPS has not been routine. The sample for the CPS is quite complex. Estimation for both the monthly labor force characteristics and for the March supplement employs complex estimation procedures that have substantial effect on the variance of the estimates. Generally, the Census Bureau has provided estimates of variance for the March supplement based on adjustments to generalized variances computed over a decade ago.

Although there has been theoretical progress on obtaining methods to estimate the mean square error (MSE) of small domain estimates (Ghosh and Rao 1994), these methods typically assume that the sampling variances are known exactly. Fay (1994) recently presented some empirical results assessing the behavior of several MSE estimators in the literature when the variances for each small domain were estimated from 10 degrees of freedom with normal data and showed considerable sensitivity of the MSE estimators to uncertainty in the variance estimates. Consequently, obtaining variance estimates as reliable as possible is an important issue in the overall research.

The multistage sample design for the CPS selects primary sampling units (PSUs) composed of one or more adjacent counties at the first stage. PSU's with large populations are sampled with certainty and are self-representing (SR), but small PSUs (termed non-self-representing or NSR) have been grouped into strata and one PSU selected randomly from each stratum. A component of the variability in the CPS estimates comes from drawing a sample of non-self-representing counties to represent the rest.

The sample design within each PSU is also complex. The majority of the CPS sample within each PSU is based on housing units enumerated in the previous census. Housing units have been grouped into segments of approximately 4 housing units each. The segments are drawn through systematic sampling. The frame is first sorted by characteristics associated with labor force participation. The variance of such a sample is typically a bit less than had segments been drawn through simple random sampling without replacement, because the sort of the frame followed by systematic sampling provides a degree of stratification of the segments.

The variance estimation strategy for total variance is best described by distinguishing variance estimation for SR and NSR PSUs. For the NSR PSUs, pairs and triplets of strata have been grouped into collapsed strata within each state. In the case of two PSUs in a collapsed stratum, a single degree of freedom for the collapsed stratum is obtained by contrasting the two PSU-level estimates using a modified half-sample technique. By using orthogonal contrasts, two degrees of freedom are obtained from each triplet. These procedures were previously applied to CPS variance estimation and for other Census Bureau surveys.

The methodology to estimate variances for SR areas is relatively novel. Some previous variance calculation efforts, including an effort based on labor force data for 1987, typically systematically divided the segments in each SR PSU into 2 half samples in an attempt to capture the effect of systematic sampling on the variances. Section 2.2 describes a new replication method employed to estimate variances for annual demographic supplements from March 1990 - March 1994 (income years 1989-1993).

Section 2.3 then describes how the effect of complex estimation methods has been taken into account.

2.2 Successive Difference Replication

Among the estimators that Wolter (1985, ch. 7) studied, two estimators, both based on squared differences between neighboring sample cases, did relatively well as general solutions. Expressed as estimators of the variance of the estimated population total

$$\hat{Y}_0 = \sum_{i=1}^{n} y_i$$

they were:

$$v_2 = (1-f)\frac{n}{2(n-1)}\sum_{i=2}^{n}(y_i - y_{i-1})^2$$

and:

$$v_3 = (1-f)\sum_{i=1}^{n/2}(y_{2i} - y_{2i-1})^2$$

where y_i, $i = 1,..., n$, represents a systematic sample from an ordered population, and f denotes the sampling fraction n/N. Estimator v_3 assumes that n is an even number. Both estimators employ squared differences of neighboring observations to estimate variation; v_3 by comparing $n/2$ distinct pairs, while v_2 compares each sample observation, except for $j = 1$ and $j = n$, to two others.

The general method of Fay (1984) could be used to provide a replication version of either variance estimator, but implementation of this approach would be difficult and requires additional work to extend the methodology appropriately to the addition of new sample units over time.

Consider instead the following modification to v_2:

$$v_{2m} = 1/2\,(1-f)\left[(y_n - y_1)^2 + \sum_{i=2}^{n}(y_i - y_{i-1})^2\right]$$

This estimator adds a comparison of the first and last sample case. In applications where the order of the sort is highly informative and y_1 and y_n are likely to be highly dissimilar, this step cannot be taken lightly. It is used here, however, to establish a link between v_2 and the successive difference replication method.

Plackett and Burman (1946) provided a method of constructing orthogonal matrices, $A = \{a_{ij}\}$ of order $4k$ such that $AA' = 4kI$, with each $a_{ij} = 1$ or -1, for most values of k up to 400 or more. In turn, most of the matrices constructed in this manner have a first row consisting entirely of 1's. Let $4k$, at least $n+2$, be the order of such a matrix, A. Then, for each replicate $r = 1,...4k$, assign to each observation y_i the replicate factor

$$f_{ir} = 1 + (2)^{-3/2} a_{i+1,r} - (2)^{-3/2} a_{i+2,r}$$

for $i < n$, and

$$f_{nr} = 1 + (2)^{-3/2} a_{n+1,r} - (2)^{-3/2} a_{2,r}$$

In turn these replicate factors define a set of replicate totals

$$\hat{Y}_r = \sum_{i=1}^{n} f_{i,r}\, y_i$$

The resulting replicate variance estimate,

$$v_{r2m} = 4(4k)^{-1}(1-f) \sum_{r=1}^{4k} (\hat{Y}_j - \hat{Y}_0)^2$$

is identical to v_{2m} since

$$\hat{Y}_r - \hat{Y}_0 = \sum_{i=1}^{n} (f_i - 1) y_i$$

$$= 2^{-3/2}\left[(y_1 - y_n) a_{2,r} + \sum_{i=1}^{n-1} (y_{i+1} - y_i) a_{i+2,r}\right]$$

$$v_{r2m} =$$

$$k^{-1}(1-f) \sum_{r=1}^{4k} 2^{-3} [(y_1 - y_n) a_{2,r} + \sum_{i=1}^{n-1} (y_{i+1} - y_i) a_{i+2,r}]^2$$

The cross-product terms for different rows of *A* vanish,

$$v_{r2m} =$$

$$k^{-1}(1-f) \sum_{r=1}^{4k} 2^{-3}\left[(y_1 - y_n)^2 a_{2,r}^2 + \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 a_{i+2,r}^2\right]$$

$$v_{r2m} = k^{-1} 2^{-3}(1-f)[((y_1 - y_n)^2 \sum_{r=1}^{4k} a_{2,r}^2$$

$$+ \sum_{i=1}^{n-1} ((y_{i+1} - y_i)^2 \sum_{r=1}^{4k} a_{i+2,r}^2)]$$

reducing the expression to v_{2m}, since

$$\sum_{r=1}^{4k} a_{i+1,r}^2 = 4k$$

for all *i*.

This result establishes the exact equivalence of v_{2m} and v_{r2m}. The intention, however, is to introduce the *successive difference replication* variance estimator, v_{r2}, obtained by defining

$$f_{ir} = 1 + (2)^{-3/2} a_{i+1,r} - (2)^{-3/2} a_{i+2,r}$$

for all *i*. Consequently, y_1 and y_n no longer share a row of the Hadamard matrix, and each is compared directly to only one other observation rather than two.

Strictly speaking, the variance estimator is not unbiased for simple random samples, but Monte Carlo studies suggest a small bias, less than 1 percent, even for relatively small *n*.

The convenience of this approach arises from the ability to assign new units to the end of the chain. For example, if a new segment is the *n*+1 to be assigned in a PSU, its replicate factor may be defined in the same manner. Now, segment *n* is compared to 2 segments, and segment *n*+1 is directly compared to only one, until segment *n*+2 is added in the future.

2.3 Representing the Effect of Complex Estimation

The effect of complex estimation on CPS estimates can be determined by implementing the complex estimator on both the original sample and each of the replicate samples. Although this has been done in specialized software previously, we implemented the March supplement weighting as a VPLX application. VPLX (Fay 1995) is a general program to implement variance estimation through replication. We have successfully replicated the most complex steps of the March CPS weighting for the supplements from March 1990 - March 1994. The implementation permits estimation of covariances of estimates across time (for any one March, approximately 50% of the sample housing units were in the sample in the previous year, making covariance across time an important consideration). For each state, we have now produced an estimated variance-covariance matrix for the six characteristics crossed by year, both for estimates of within PSU variance (ignoring the effect of the first stage selection of PSUs) and total variance.

2.4 Remaining Variance Research

Although the successive difference variance estimation produces, we believe, relatively stable estimates of the within-PSU variances for most states, estimation of total variance remains less stable because of the use of a collapsed stratum method to estimate variance for the non-self-representing PSUs. Otto and Bell have begun to model the relationship between the estimated within variance and the estimated total variance in each state.

3. Small Domain Estimation Research

The Census Bureau has also produced annual estimates of median income for 4-person families by state based on a model combining CPS and administrative estimates (Fay, Nelson, and Litow 1993). The model for 4-person family medians incorporates multivariate features potentially useful for this application. This section describes preliminary calculations, based on census data, indicating that similar procedures, adapted to estimation of household medians and per capita income (PCI), appear quite promising to provide an initial model.

The methodology has been applied for 4-person families separately for each year, *t*, in the series. (For simplicity, the implicit subscript, *t*, is not shown in the following, except where necessary to avoid confusion.) As summarized by Fay, Nelson, and Litow (1993), the primary elements of the current methodology for 4-person medians are:

1) For each state, *s*, (and the District of Columbia), a direct sample estimate, $\hat{Y}_{s4}$, of the median income for 4-person families is estimated from the CPS.
2) Similarly, median incomes for 3- and 5-person families, $\hat{Y}_{s3}$ and $\hat{Y}_{s5}$, are estimated as well. For each state, the weighted combination of the two medians,

$$\hat{Y}_{sc} = .75\,\hat{Y}_{s3} + .25\,\hat{Y}_{s5}$$

is computed. The weights, .75 and .25, are approximately proportional to the respective sample sizes, in other words, there are roughly 3 times as many 3-person families as 5-person families.
3) Regressions are fitted to $\hat{Y}_{s4}$ and $\hat{Y}_{sc}$, with separate predictors and corresponding coefficients for each of these two variables. The regressions produce fitted values, $\hat{Y}_{(REG),s4}$ and $\hat{Y}_{(REG),sc}$. The regression model for medians of 4-person families employs 3 predictor variables:
   a) $X_{s41} = 1$, to correspond to a constant term in the model.
   b) $X_{s42} = (BEA_{st}/BEA_{sb})\,Y_{(CEN),s4}$, where BEA_{st} represents Bureau of Economic Analysis (BEA) PCI for the same income year, *t*, as $\hat{Y}_{s4}$, and BEA_{sb} and $Y_{(CEN),s4}$ represent BEA PCI and census median income for 4-person families, respectively, for the same base income year, *b*, of the previous census. This predictor variable thus represents the census median adjusted by the proportional growth in BEA PCI since the previous census.
   c) $X_{s43} = Y_{(CEN),s4}$, that is, median incomes from the previous census.

   The regression model for the weighted average, $\hat{Y}_{sc}$, uses an analogous set of variables, $X_{sc1} = 1$, $X_{sc2} = (BEA_{st}/BEA_{sb})\,Y_{(CEN),sc}$, and $X_{sc3} = Y_{(CEN),sc}$.
4) A composite estimate, $\hat{Y}_{(COMP),s4}$, is formed from $\hat{Y}_{s4}$, $\hat{Y}_{sc}$, $\hat{Y}_{(REG),s4}$, and $\hat{Y}_{(REG),sc}$. The combination of the direct sample estimate for 4-person families, $\hat{Y}_{s4}$, with the regression estimate for 4-person families, $\hat{Y}_{(REG),s4}$, is a feature that has appeared in other small domain estimation models based on empirical Bayes procedures. The methodology is in fact multivariate, in using further information present in $\hat{Y}_{sc}$ and $\hat{Y}_{(REG),sc}$ to estimate medians for 4-person families.

Note that this procedure may be generalized to estimate more than 2 variables simultaneously. In the application to 4-person family medians, the weighted average, $\hat{Y}_{sc}$, is included to improve the estimation of the 4-person family median. Generally, we would expect to apply the multivariate model to household medians, per capita income, and the poverty characteristics of interest, without introducing additional dependent variables.

Data from the 1980 and 1990 censuses indicate that this model may be directly adapted for per capita income and household medians simply by substituting the analogous variables into steps 1), 2), 3)b, and 3)c. With the 1990 census values as dependent variables at steps 1) and 2), and with the 1980 census values in 3)b and 3)c, the adjusted R^2 is .935 for household medians and .979 for per capita income. The success of the model is suggested by figures 1 and 2. The root mean square errors for the two models are $1148 and $370, respectively, considerably smaller than the standard errors of the direct CPS estimates in most states. Thus, we expect a composite formed from the CPS and the model to have considerably greater stability across time and lower MSE than the direct estimate alone.

The results for median income and PCI are each univariate. The correlation between error terms of the two models is relatively modest, so that we do not anticipate the multivariate gains as for the 4-person family median problem. We do expect, however, to examine the multivariate approach for estimating the four components of the poverty population.

The form of the model does not lend itself so immediately to the estimation of poverty, however. Work of John Coder and Paul Siegel suggests that proportion of

low income tax returns and food stamp participation may be better indicators of state poverty rates than changes in BEA PCI. These issues will be investigated further.

[1] This article reports results of research undertaken by staff members of the Census Bureau. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. The authors thank Lynn Weidman for helpful comments.

References

Fay, R. E. (1984), "Some Properties of Estimates of Variance Based on Replication Methods," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 495-500.

_____ (1994), "The Role of Design Based Variances and Covariances in Small Domain Estimation," presented at the Seminar on New Directions in Statistical Methodology, Council of Professional Associations in Federal Statistics, Bethesda, MD, May 25-26, 1994.

_____ (1995), "VPLX: Variance Estimates for Complex Samples, Program Documentation," unpublished Census Bureau report.

Fay, R. E., Nelson, C. T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in Schaible, W. and Gonzalez, M. E. (eds.), *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC., pp. 9-1 - 9-17.

Ghosh, M., and Rao, J.N.K. (1994), "Small Area Estimation: An Appraisal (with discussion)," *Statistical Science*, **9**, 55-93.

Long, J. (1993), "Postcensal Population Estimates: States, Counties and Places," in Schaible, W. and Gonzalez, M. E. (eds.), *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC., pp. 3-1 - 3-30.

Plackett, R. L. and Burman, J. P. (1946), "The Design of Optimal Multifactorial Experiments," *Biometrika*, **33**, 305-325.

Schaible, W.L. and Gonzalez, M.A., eds. (1993), "Indirect Estimators in Federal Programs," *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

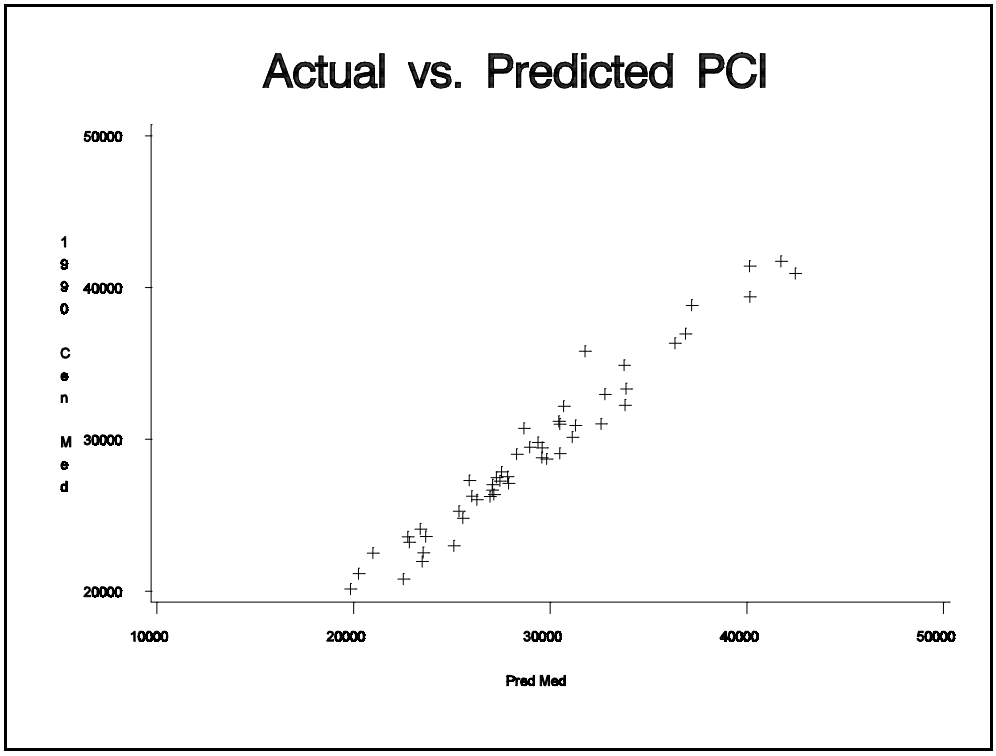

Figure 1. Comparison of Predicted 1989 Census Medians from the BEA equation and 1980 census data to the values from the 1990 Census.

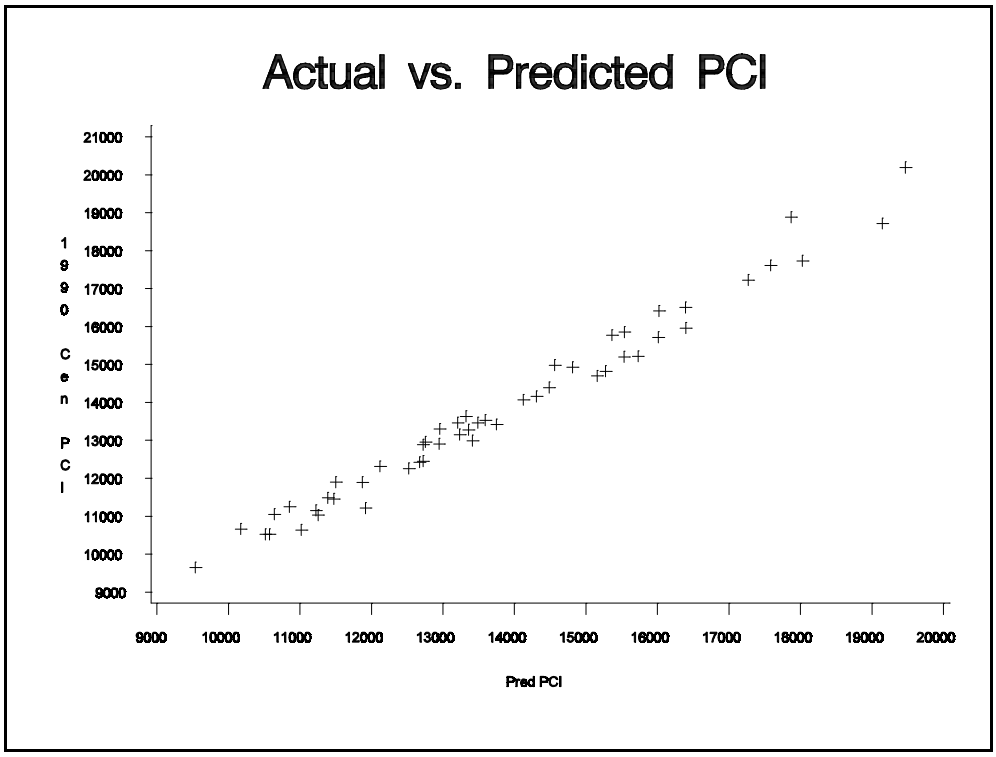

Figure 2. Comparison of Predicted 1989 Census Per Capita Income from the BEA equation and 1980 census data to the values from the 1990 Census.