SMALL DOMAIN METHODOLOGY FOR ESTIMATING INCOME AND POVERTY CHARACTERISTICS FOR STATES IN 1993

Robert E. Fay and George F. Train
U.S. Bureau of the Census

Prepared for presentation at the Joint Statistical Meetings, Anaheim, CA, August 10-14, 1997.

SMALL DOMAIN METHODOLOGY FOR ESTIMATING INCOME AND POVERTY CHARACTERISTICS FOR STATES IN 1993

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 20233-9001

Key Words: Small area estimation, Empirical Bayes

1. INTRODUCTION

This paper reports methodology underlying estimates for income year 1993 of income and poverty by state produced as part of the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) project. The paper has been prepared for a session devoted to this project, following an example set two years ago in which Siegel (1995) reviewed issues and strategies for county estimates, Fay and Train (1995) described methodology to produce direct variance estimates through replication for states and presented preliminary modeling results for per capita and median income at the state level, Otto and Bell (1995) described modeling of the direct variance estimates at the state level, and Cohen (1995) provided an initial reaction to these preliminary results.

The original objectives of the SAIPE project were to produce estimates for states and counties of:

1. total persons in poverty,
2. related children age 0-4 in poverty,
3. related children age 5-17 in poverty,
4. persons age 65 and over in poverty,
5. median household income, and
6. per capital income.

Initial plans were to release state and county estimates of these six characteristics every two years beginning with estimates for income year 1993.

On March 26, 1997, the Census Bureau released state and county estimates for 1993 for three of these characteristics: total poverty, related children 5-17 in poverty, and median income. The estimates and additional documentation of the project are available at http://www.census.gov/hhes/www/saipe.html or through the Census Bureau's web site, http://www.census.gov. Although state estimates have been produced for the remaining three components, county estimates are not ready for release. There has been a recent redefinition of the objectives. The Census Bureau now plans to continue to produce estimates for the three characteristics already released, plus total poor age 0-17 and median household income at both the state and county level, and estimates of poor age 0-4 at the state level only.

Estimates of poverty for related children age 5-17 have thus far attracted the most attention. Reauthorization of the Elementary and Secondary Education Act (the Improving America's School Act of 1994) directed the Secretary of Education to employ updated estimates of poor children age 5-17 more recent than the 1990 census if suitable estimates could be produced by the Census Bureau. The legislation further mandated that the Secretary of Education consult with a special panel of the National Academy of Sciences convened for the purpose of assessing the accuracy of the estimates. Because the allocation formulas are specified at the county and school district levels, the panel focused its attention primarily toward the county estimates. The panel released both an Executive Summary (available at http://www2.nas.edu/new/2146.html) and more recently an Interim Report (Citro, Cohen, Kalton, and West 1997). The report was generally critical of the 1993 estimates and could not endorse their direct use, recommending instead that the estimated 1993 poverty rates be averaged with the 1990 census rates. This recommendation has been implemented for Title I.

This paper summarizes the methodology for the 1993 state estimates. It abstracts from a draft report (Fay 1997) providing further detail. The primary focus of the paper will be on the estimates of related children age 5-17 also. The paper will also comment on how estimates for this age group exemplify the methodology for other ages and summarize the estimation of median income.

The state estimates combine information from the Current Population Survey (CPS) for the target income year, information from the previous census, and auxiliary data through a now relatively well-studied empirical Bayes procedure (Fay and Herriot 1979, Ghosh and Rao 1994). The next section summarizes features of the CPS and census estimates of poverty and the auxiliary data used in the models. The third section details the estimator. The fourth section summarizes estimation of median income. The concluding section summarizes the status of current research and issues requiring further study.

2. Sources of Data

2.1 Estimates of Poverty. The March Supplement to the CPS provides the official estimates of income and poverty nationally. Although the sample size of the CPS is sufficient for effective national estimates, generally the size is too small to permit direct estimation of poverty statistics at the state or substate level (Fay 1997). The census data has been the only source for such geographic detail. Comparison of Fig 1a to 1b illustrates that state values from the previous census can be as effective a predictor as the direct CPS state estimates.

Although income and poverty concepts are essentially identical in the CPS and census, there are some differences that affect the comparability of estimates from these two sources (Fay 1997).

Both the CPS and census data may be disaggregated by age groupings. A total of six components of the poverty population are considered: 1) related poor children 0-4, 2) related poor children 5-17, 3) poor 18-64, 4) poor 65+, 5) total poor 0-4, and 6) total poor 5-17. (Related poor children include only children related to the head who are not the head or spouse of head. Total poor are all poor in these age groups, regardless of relationship.) The estimation strategy produced estimates of proportion poor for each of these age components, enabling totals to be built up from pieces.

2.2 Auxiliary Data. The Census Bureau employs extracts from IRS tax return files for a number of statistical purposes, including postcensal population estimation. The information available to the Census Bureau from each return includes the number of exemptions, the number of child exemptions, the number of exemptions for 65+, and the adjusted gross income (AGI).

Under some circumstances, children can file returns and be claimed on a parent's single or joint return as well. A question on the tax return form asks if this is the case, and this information is also available to the Census Bureau. For statistical purposes, the child's own return is excluded from analysis, since the family circumstances of the child are better represented by the parent's return.

Although the definition of census income is different from AGI reported on IRS returns, defining IRS poor persons by applying the low income cutoffs for census poverty to IRS AGI and number of exemptions produces an effective auxiliary measure for predicting census poverty status. The number of poor child exemptions as a fraction of all child exemptions is used in the model for related children 5-17 in poverty. Families filing returns but with low income in many cases may represent the "working poor," who have limited economic resources but who derive some or all of their income from earnings rather than only benefits.

A second variable in the model complements the IRS poor, namely, the number of persons age 0-64 not on IRS returns, estimated as the difference between the Census Bureau's postcensal estimate for this age group and the number of exemptions for 0-64. This group is indicative of the size of the population dependent on transfer payments exclusively.

A third variable in the model is the proportion of the population receiving food stamps. This group partially overlaps with the first two, since both the working poor and those receiving transfer payments are eligible. Fay (1997) further discusses characteristics of the auxiliary data.

3. Estimation of Poverty Proportions

3.1 Form of the model. For the 1980 and 1990 censuses separately, a cross-sectional model was fitted using the available predictors. The form of the regression model was:

$$\begin{aligned} \textit{Census \%poverty} &= b_0 + b_1(\textit{IRS \%poor}) \\ &+ b_2(\%\ \textit{IRS nonfilers}) + b_3(\%\ \textit{fs}) \end{aligned}$$

The model is cross-sectional in the sense of only using data pertaining to a single income year. The fit of this model was relatively good in both census years.

Nonetheless, there was some observed correlation between the residuals from the fits at the two censuses. The residuals from the cross-sectional model were

$$\begin{aligned} \textit{Census resid.} &= \textit{Census \%poverty} - \\ &[b_0 + b_1(\textit{IRS \%poor}) \\ &+ b_2(\%\ \textit{IRS nonfilers}) \\ &+ b_3(\%\ \textit{fs})] \end{aligned}$$

Addition of the 1979 residual to the model for 1989 poverty substantially improved the fit.

For 5-17, for example, the coefficient on the residual was .56 with a standard error computed under ordinary least squares (OLS) assumptions of .10. Thus, the longitudinal equation for 1989 was

$$\begin{aligned} \textit{Census \%poverty}, 89 &= b_0 \\ &+ b_1(\textit{IRS \%poor}, 89) \\ &+ b_2(\%\ \textit{IRS nonfilers}, 89) \\ &+ b_3(\%\ \textit{fs}, 89) + b_4\ (79\ \textit{resid}) \end{aligned} \qquad (1)$$

Fig. 1c compares the fit of this model to the census. The comparison, of course, is optimistic since the coefficients of the model are determined from the same census data to which the fit is compared.

The final step of developing the model is replacement of the poverty rates from the decennial census with sample estimates from the CPS in order to produce current estimates. Unlike the fit to the census estimates, which have negligible sampling error at the state level, the CPS model must distinguish sampling error from model error. The model-based procedure developed for the SAIPE can be viewed as a form of empirical Bayes estimation, or as an empirical best linear unbiased

predictor (EBLUP) under a mixed linear model (Ghosh and Rao 1994).

The small domain estimator is based on the model:

$$\hat{\boldsymbol{p}}_{(CPS)} = \boldsymbol{X}\beta + \boldsymbol{b} + \boldsymbol{e} \qquad (2)$$

where β represents a vector of the regression coefficients, which are fixed effects in the model, $\boldsymbol{b}$ represents a column vector of random effects denoting the departure of the individual true values from the regression predictions, and $\boldsymbol{e}$ denotes a column vector of CPS sampling errors.

The random effects, $\boldsymbol{b}$, are assumed normal and to have mean **0** and a diagonal covariance matrix.

$$\boldsymbol{A}^* = \begin{bmatrix} \boldsymbol{A} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \boldsymbol{A} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A} & \ldots \\ . & . & . & \ldots \end{bmatrix} \qquad (3)$$

The CPS sampling errors $\boldsymbol{e}$ are assumed uncorrelated with the random effects, normal with mean **0** and diagonal covariance matrix

$$\boldsymbol{D}^* = \begin{bmatrix} \boldsymbol{D}_1 & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \boldsymbol{D}_2 & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D}_3 & \ldots \\ . & . & . & \ldots \end{bmatrix} \qquad (4)$$

Given both $\boldsymbol{A}^*$ and $\boldsymbol{D}^*$, the best linear unbiased estimate (BLUE) of β is

$$\hat{\beta} = (\boldsymbol{X}'(\boldsymbol{D}^* + \boldsymbol{A}^*)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{D}^* + \boldsymbol{A}^*)^{-1}\hat{\boldsymbol{p}}_{(CPS)}, \qquad (5)$$

and the BLUE of the expected values of the CPS estimates,

$$E_p(\hat{\boldsymbol{p}}_{(CPS)}) = \boldsymbol{X}\beta + \boldsymbol{b}, \qquad (6)$$

is

$$\hat{\boldsymbol{p}}_{(CPS,\ comp)} = \boldsymbol{X}\hat{\beta} + \boldsymbol{A}^*(\boldsymbol{D}^* + \boldsymbol{A}^*)^{-1}(\hat{\boldsymbol{p}}_{(CPS)} - \boldsymbol{X}\hat{\beta}) \qquad (7)$$

$\boldsymbol{A}$. in (3) is not known, however. Estimation of $\boldsymbol{A}$., followed by use of its estimate in (5) and (6), results in an empirical BLUP (EBLUP) estimator. MLE was used to estimate $\boldsymbol{A}$.

In the normal model, the components of (4) are assumed fixed and known. Because the model involves proportions, however, the mean and variance are linked. The approach taken to this slightly nonstandard version was to modify the components of (4) by an adjustment to reflect the underlying expected values.

The variance model of Bell and Otto (1995) was based on original proportions p_{0st} and provided estimates of variance $\hat{D}(p_{0st})$. Their findings were generalized to arbitrary p_{0st} by assuming a constant design effect under a binomial model:

$$\hat{D}(p_{st}) = \frac{p_{st}(1 - p_{st})}{p_{0st}(1 - p_{0st})}\hat{D}(p_{0st}) \qquad (8)$$

The estimation was iterative. First, *eq.* (8) was applied to the CPS estimates and the resulting variances substituted into (4). After MLE of $\boldsymbol{A}$, the resulting regression predictions $\boldsymbol{X}\hat{\beta}$ were then used in (8). A total of 6 iterations were performed; examination of the intermediate results suggested that virtually all of the effect of the iteration had occurred by the end of the first 3 cycles. MLE estimate of $\boldsymbol{A}$ was 0 for related children 5-17. Fig. 1d shows the performance of the resulting estimates, which approximates the performance of the fit of the model to the census values, Fig. 1c, and is clearly better than the previous census, 1a, or the direct CPS estimates, 1b.

In application to the 1994 CPS, the estimated value of $\boldsymbol{A}$ for related children 5-17 was .37 when the dependent variable was expressed on the percent scale (equivalent to .000037 as a proportion).

Estimates of poverty for 0-4, 18-64 and 65+ were obtained similarly. The set of auxiliary predictors for 0-4 and 18-64 was quite similar to those for 5-17; those for 65+ were somewhat less similar, substituting % receiving Supplemental Security Income for % Food Stamps, for example. Evaluations of the performance of the models for 0-4, 5-17 and total poverty were all encouraging, but clear evidence of improvement over previous census distributions for the 65+ was lacking (Fay 1997). Table 1 gives the estimated coefficients.

4. Estimation of Median Income

Sample estimates of state median incomes from the CPS are more stable relatively than poverty estimates, and for a number of years the Census Bureau has reported this variable in an annual publication. Nonetheless, preliminary research reported by Fay and Train (1995) suggested the potential improvement from a similar

modeling approach. The model for median income employed a smaller set of predictors: a constant term, median income in the previous census, and a predicted value of median income based on inflating the census median by the relative growth in AGI per return from IRS.

5. Some Remaining Research Issues

The SAIPE project is ongoing. Because of the potential use of the 1993 estimates for a variety of purposes, documentation and review of the 1993 experience is important on its own merit. The state methodology may be modified in subsequent years by current research. For example, Bell and Otto (1997) discussed the potential application of time series methods to the state estimation problem, but the determining factor in not applying these methods was the absence of key IRS variables for most of the years. As this situation has changed, the question will require careful examination.

IRS variables for additional years have recently become available. Fitting of the state models for these years would provide additional evidence on the performance and stability of the state model.

The NAS report (Citro, Cohen, Kalton, and West 1997) raises a number of questions about the performance of the estimation procedure, including contrasting strategies employed by the state and county models. Further study is warrented here, not only for the important issue of Title I allocations, but as an interesting case study in small area estimation.

[1] This paper 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 Nanak Chand and Mary Ann Cochran for helpful comments.

REFERENCES

Bell, W.R. and Otto, M.C. (1997), "Bayesian Inference about Poverty and Income for States," to be presented at the Annual Meetings of the American Statistical Association, Anaheim, CA, Aug. 10-14, 1997.

Citro, C.F., Cohen, M.L., Kalton, G., and West, K.K., eds. (1997) "Small-Area Estimates of School-Age Children in Poverty: Interim Report I-Evaluation of 1993 County Estimates for Title I Allocations," National Academy Press, Washington, DC.

Cohen, M.L. (1995), "Discussion," *Proceedings of the Section on Government Statistics,* American Statistical Association, Alexandria, VA, pp. 172-174.

Fay, R.E. (1997) "Estimates of Poverty and Income by State for Income Year 1993," draft Census Bureau report dated Jan. 30, 1997.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association,* **78**, 269-277.

Fay, R.E. and Train, G.F. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *Proceedings of the Section on Government Statistics,* American Statistical Association, Alexandria, VA, pp. 154-159.

Fisher, R.C. and Siegel, P.M. (1997), "Methods Used for Small Area Income and Poverty Estimation," to be presented at the Annual Meetings of the American Statistical Association, Anaheim, CA, Aug. 10-14, 1997.

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

Otto, M.C. and Bell, W.R. (1995), "Sampling Error Modeling of Poverty and Income Statistics for States," *Proceedings of the Section on Government Statistics,* American Statistical Association, Alexandria, VA, pp. 160-165.

Siegel, P.M. (1995), "Developing Postcensal Income and Poverty Estimates for All US Counties," *Proceedings of the Section on Government Statistics,* American Statistical Association, Alexandria, VA, pp. 166-171.

Siegel, P.M., and Fisher, R.C. (1997), "1993 Income and Poverty for US Counties: Some Methodology of and Results from the Census Bureau Small Area Income and Poverty Estimates," to be presented at the Annual Meetings of the American Statistical Association, Anaheim, CA, Aug. 10-14, 1997.

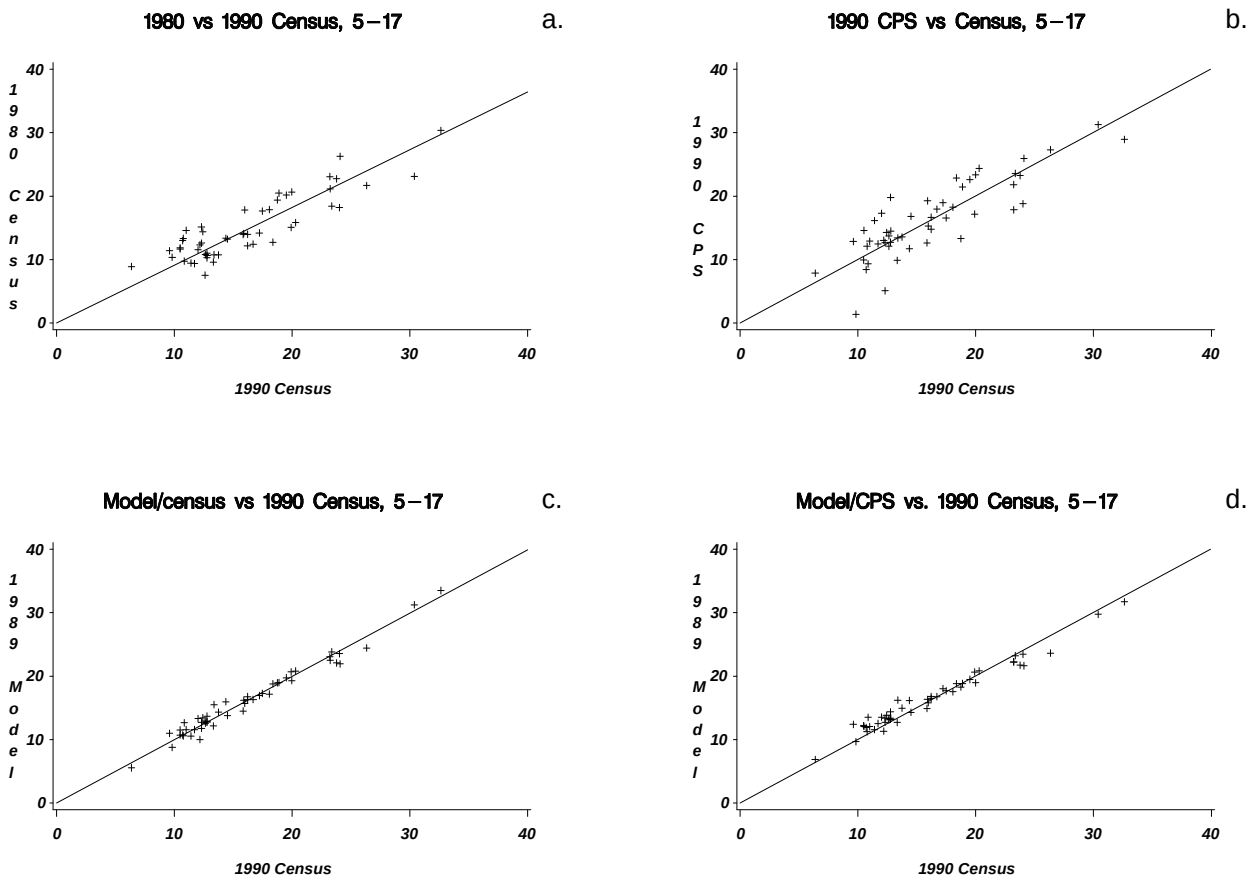

Fig. 1. Comparison of alternative estimates of poverty rates in 1989 (y-axis) and with 1990 census results (x-axis) for states, for related children 5-17. There is a statistical association between the 1980 and 1990 census values (a.). Direct survey estimates from the CPS (b.) appear somewhat less related to the 1990 census values than the 1980 census, showing that the sampling variability of the direct estimates is so large as to make the previous census values a better indication of the relative distributions among the states. When the regression model is fitted to the 1990 census values, the fits are quite good (c.). Fitting the regression to the 1990 CPS appears to lead to a modest loss of prediction relative to (c.) but the results are clearly better than (a.) or (b.). The line shown each case is the OLS regression through the origin.

Table 1. Coefficient estimates for the regression fit to 1993 poverty rates from the 1994 CPS. For age 65+, the equation employs the poverty rate from the 1990 census instead of the 1989 cross-sectional residual. Standard errors in parenthesis reflect both CPS sampling error and the estimated model variance.

	Rel ch 0-4	0-4	Rel ch 5-17	5-17
1989 residual	1.06 (.47)	.92 (.47)	1.36 (.39)	1.23 (.41)
% poor exemptions 0-64	.65 (.24)	.72 (.23)		
% poor child exemptions			.31 (.11)	.34 (.11)
% nonfilers 0-64	.59 (.20)	.65 (.20)	.52 (.13)	.55 (.14)
% food stamps	.97 (.33)	.90 (.32)	.98 (.22)	.99 (.23)
Constant	-2.05 (3.31)	-2.02 (3.24)	-3.39 (1.98)	-3.26 (2.08)
	18-64		65+	
1989 residual	.81 (.30)			
1989 census %			.80 (.14)	
% poor exemptions 0-64	.54 (.08)			
% poor exemptions 65+			-.12 (.29)	
% nonfilers 0-64	.27 (.08)			
% nonfilers 65+			.02 (.09)	
% food stamps	.37 (.11)			
% SSI 65+			-.07 (.14)	
Constant	-2.59 (1.17)		2.60 (4.40)	