ESTIMATION OF MEDIAN INCOME FOR 4-PERSON FAMILIES BY STATE

Robert E. Fay, Charles T. Nelson[1], U.S. Bureau of the Census
Leon Litow, U.S. Department of Health and Human Services
Robert E. Fay, U.S. Bureau of the Census, Washington, DC 20233-4001

KEY WORDS: *Small domain estimation, empirical Bayes, Current Population Survey*

Abstract For purposes of comparison to other small domain estimation methodologies in actual use in federal statistical programs, this paper describes a small domain estimation program at the U.S. Census Bureau to estimate median family incomes by state annually, using data from the decennial censuses, the Current Population Survey, and estimates produced by the Bureau of Economic Analysis. The current procedure is essentially a multivariate empirical Bayes estimator. The 1990 census affords the opportunity to assess the performance of the model against actual census results. Although the analysis stresses areas for possible improvement, the comparisons to the 1990 census are favorable and suggest the continued usefulness of this approach.

1. INTRODUCTION

This paper is an outgrowth of an effort by the Subcommittee on Small Area Estimation, chaired by Wesley Schaible, of the Federal Committee on Statistical Methodology, chaired by Maria E. Gonzalez, to document indirect estimators published by federal statistical agencies. The paper extracts highlights from a chapter (Fay, Nelson, and Litow 1993) of the subcommittee's report.

Starting with income year 1974, the U.S. Census Bureau has computed model-based estimates of median annual income for 4-person families by state using data from the decennial censuses, the Current Population Survey (CPS), and estimates of per capita income (PCI) from the Bureau of Economic Analysis (BEA). Originally, these estimates were used in determining eligibility for the former Title XX Programs of the Social Security Act; the estimates are now used in the administration of the Low Income Home Energy Assistance Program. Fay, Nelson, and Litow (1993) further describe the programmatic use of the estimates.

In addition to their programmatic use, the estimates represent the only intercensal state-specific family income estimates produced by the Census Bureau. Consequently, these estimates have been of interest to a number of general data users. Until the recent publication of the historical series in U.S. Bureau of the Census (1991), however, the estimates did not appear in a regular publication series of the Census Bureau.

2. GENERAL APPROACH

Throughout this period the methodology has relied on three sources:

1. Estimates of median family income by state from the decennial censuses. Since the census asks income during the previous year, the census medians pertain to income years 1969, 1979, etc. Although the estimates are based on the long-form sample, the size of this sample provides estimates with virtually negligible sampling errors at the state level every 10 years.
2. Sample estimates of median income by family size by state from the CPS. Although the CPS estimates are available annually, their direct use is limited by substantial sampling variability due to the size of the CPS sample.
3. Annual estimates of PCI from BEA. These estimates, based on aggregate statistics on components of income from administrative series, have negligible sampling error. The PCI estimates are measures of average income per person, however, and so are only indirectly linked to median income for families. There are also important conceptual differences between these estimates and census income. Bailey, Hazen, and Zobronsky (1993) further describe these estimates.

Fay, Nelson, and Litow (1993) describe these three sources in further detail, and cite additional references.

Before outlining the elements of the methodology, we first compare the estimates for income year 1989, based on the March 1990 CPS and published in March, 1991, with medians for 4-person families obtained by special tabulation of the 1990 census in September, 1992. Figure 1 shows the geographic distribution of the true increase in median income for 4-person families during the decade, since the 1980 census.

Figure 1 indicates that the greatest relative increase during the decade in the median income of 4-person families occurred in the Northeast region, where most states more than doubled their medians, according to the census. Other areas of active growth include additional states in the East and South Atlantic, and Tennessee, Minnesota, California, and Hawaii. Figure 1 also shows that median income in some areas of the country has grown considerably more slowly.

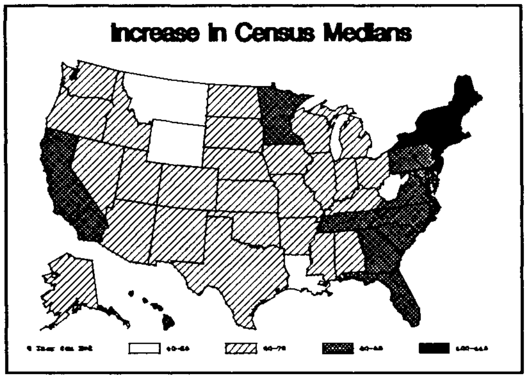

Figure 1 Percent increase in census median income for 4-person families, 1979-1989.

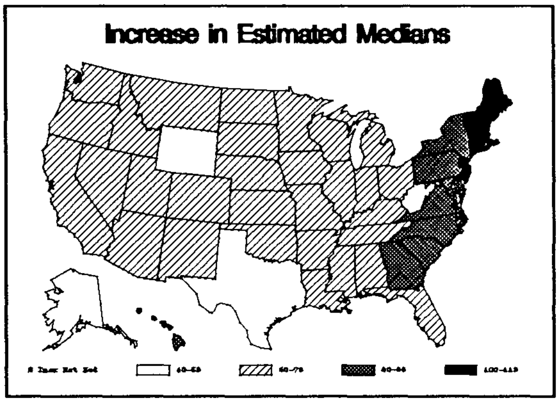

Figure 2 Estimated increase in median incomes of 4-person families, 1979-1989, comparing the estimated 1989 medians to the 1980 census values for 1979 medians.

Figure 2 presents the estimated increase since the 1980 census according to the model. Although there are some differences between the census and the model predictions, the comparison of the two maps shows that the model is successful in capturing most real sources of change in median family income. Some of the states are not classified into the same grouping in Figures 1 and 2, but, in each case, the difference is by at most one category. For example, states estimated to be among the fastest growing group were either in that category or the next one down, and so forth.

Figure 3 shows the geographic distribution of the key predictor variable, the change in estimated BEA per capita income. Note that the scale of percent income growth is shifted on this third map compared to the other two; in general, the proportional increase in per capita personal income outstripped the increase in the median income of 4-person families during the decade. With the rescaling, however, the BEA income figures are quite successful predictors of the corresponding change in the median income of 4-person families at the state level.

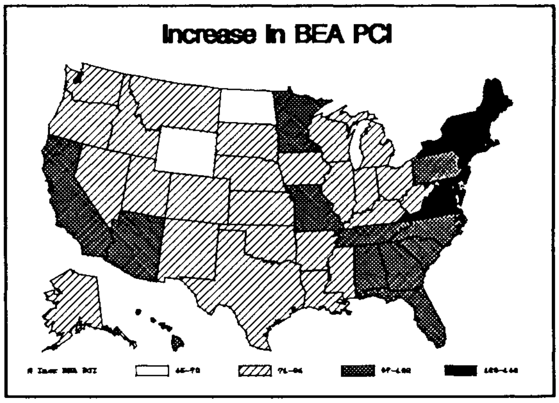

Figure 3 Percent increase in BEA PCI for 1979-1989.

Figures 4 and 5 illustrate additional features of the performance of the model. In both cases, a regression line from the simple linear regression appears as an aid in assessing fit, although each line is not formally included in the model.

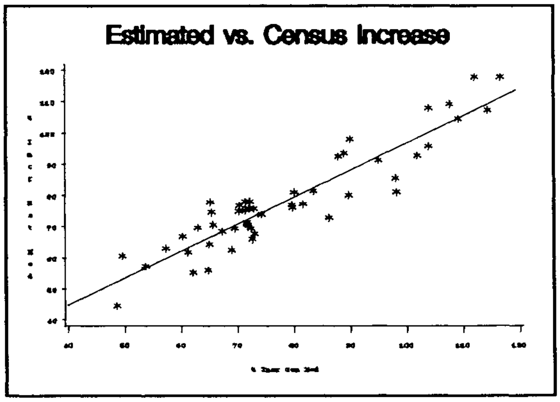

Figure 4 Percent increase in predicted median incomes compared to the 1990 census, both relative to 1980 census values of 1979 medians.

Figure 4 compares the estimated change with the actual change in median incomes, according to the census. Again, the predictions are not perfect but, nonetheless, appear to capture most of the variation among states in the increase of median income. Figure 5 shows that the relationship between increase in the census median income is essentially linear over the entire spectrum. As previously indicated by the scaling Figure 3, Figure 5 provides further evidence of somewhat greater dispersion in the change in the BEA estimates than in the census medians.

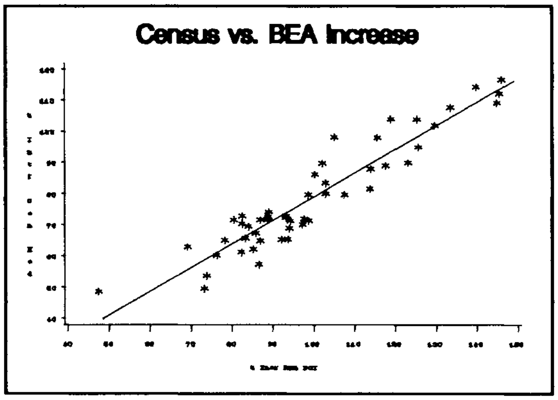

Figure 5 Percent increase in census medians, 1979-1989, compared to the percent increase in BEA PCI.

The current methodology has been in place since income year 1984, although with minor refinements over this period of time. The methodology is applied 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. Section 5 will discuss the possibility of alternative forms more attuned to the longitudinal nature of the problem.) The primary elements of the current methodology 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. The medians actually are obtained by linear interpolation using tabulated income categorized into intervals of \$2,500. (The census medians were also estimated by interpolation of categorized data.)
2) Similarly, sample estimates of 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 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 increase 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 analogous 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.

Figure 5 motivates the inclusion of both $X_{s42} - (BEA_{st}/BEA_{sb})\,Y_{(CEN),s4}$, and $X_{s43} - Y_{(CEN),s4}$, as predictors in the model. Given the consistent linear relationship between proportional change in BEA PCI and change in the census median, the first of these two terms is the most obvious single expression of this relationship. The second of the two complements the first: without the second, the regression would be satisfactory only if the slope in Figure 5 were 1.0. In fact, the slope is somewhat less than 1. Inclusion of both predictors allows, in effect, the slope of the regression line in Figure 5 to be estimated from the CPS data.

A key feature of the model is the multivariate combination of estimation of the target variable of interest, median income for 4-person families, along with an auxiliary variable, the combined 3- and 5-person family medians, even though the auxiliary variable is not itself a subject of interest. In fact, the purpose of the multivariate approach is to realize additional gains in the estimation of 4-person family medians. Fuller and Harter (1987) and Fay (1987) motivate the possible advantages of the multivariate approach for problems of this sort.

3. ESTIMATOR

3.1 General Features of the Current Estimator.

The last two steps, 3) and 4), of the current procedure can be represented more completely in matrix notation. Let $\hat{\boldsymbol{Y}}$ represent a 102-element column vector formed from the CPS sample medians in each state, i.e.,

$$\hat{\boldsymbol{Y}} = (\hat{Y}_{14}, \hat{Y}_{1c}, \hat{Y}_{24}, \hat{Y}_{2c}, \ldots, \hat{Y}_{51,4}, \hat{Y}_{51,c})'$$

Let $\boldsymbol{X}_{s4} = (X_{s41}, X_{s42}, X_{s43})$ denote a row vector with the 3 predictor variables for the 4-person family median in state s, and $\boldsymbol{X}_{sc} = (X_{sc1}, X_{sc2}, X_{sc3})$ the predictors for the weighted combined variable, $\hat{Y}_{sc}$, and let

$$\boldsymbol{X} = \begin{pmatrix} X_{14} & 0 \\ 0 & X_{1c} \\ X_{24} & 0 \\ 0 & X_{2c} \\ \cdots & \cdots \end{pmatrix}$$

be a 102 by 6 matrix containing all the predictors. The small domain estimator is based on the model:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}$$

where $\boldsymbol{\beta}$ represents a 1 by 6 row vector of the regression coefficients, $\boldsymbol{b}$ represents a 102 by 1 column vector of random effects denoting the departure of the individual true medians from the regression predictions, and $\boldsymbol{e}$ denotes a 102 by 1 column vector of sampling errors. The covariance matrix of $\boldsymbol{b}$ is assumed to be in a block diagonal form:

$$\boldsymbol{A}^* = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & \ldots \\ \boldsymbol{0} & \boldsymbol{A} & \boldsymbol{0} & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A} & \ldots \\ . & . & . & \ldots \end{bmatrix}$$

where $\boldsymbol{A}$ represents a 2 by 2 covariance matrix for the model errors. Note that the model assumes that the distribution of the random effects, $\boldsymbol{b}$, is identical in each state and uncorrelated between states. The sampling covariances of $\boldsymbol{e}$ are assumed to be of the form

$$\boldsymbol{D}^* = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{0} & \ldots \\ \boldsymbol{0} & \boldsymbol{D}_2 & \boldsymbol{0} & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_3 & \ldots \\ . & . & . & \ldots \end{bmatrix}$$

where $\boldsymbol{D}_s$ is a 2 by 2 covariance matrix for the sampling errors for state s. Consequently, unlike the assumption that the random effects have the same distribution in each state, $\boldsymbol{D}_s$ is allowed to vary over states.

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{D}^*+\boldsymbol{A}^*)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{D}^*+\boldsymbol{A}^*)^{-1}\hat{\boldsymbol{Y}},$$

and the BLUE of the true medians,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b}$$

is

$$\hat{\boldsymbol{Y}}_{(COMP)} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{A}^*(\boldsymbol{D}^* + \boldsymbol{A}^*)^{-1}(\hat{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.1}$$

Because of the block diagonal forms of $\boldsymbol{A}^*$ and$\boldsymbol{D}^*$, this estimator forms a weighted average of the sample and regression estimates within each state for both the 4-person and combined family medians, as described earlier at 4).

Estimator (3.1) has been employed for income years 1984-1991, but refinements have entered in the estimation of the sampling errors, $\boldsymbol{D}_t$, and model errors, $\boldsymbol{A}$. Fay, Nelson, and Litow (1993) describe the approach used to estimate them.

3.2 Estimation of the Model Errors As noted earlier, $\boldsymbol{A}^*$ is assumed to have a block diagonal form with identical 2 by 2 covariances, $\boldsymbol{A}$, on the diagonal. In other words, the model errors in each state are assumed to have an identical distribution.

Empirical Bayes estimates characteristically estimate components of model error, such as $\boldsymbol{A}$, directly from the data. A typical experience, however, is that such estimates are themselves subject to considerable variation, which in turn increases the variance of estimators such as (3.1) compared to the variance of (3.1), if $\boldsymbol{A}$ were known.

By fitting the regression model based on the 1970 census medians and changes in BEA PCI between 1969 and 1979 to the 1980 census values, an estimate of model variance may be obtained. The results were:

$$\boldsymbol{A}_{(CEN)} = \begin{pmatrix} 0.308 & 0.263 \\ 0.263 & 0.286 \end{pmatrix} \times 10^6$$

Until income year 1989, $\boldsymbol{A}$ has been estimated by projecting $\boldsymbol{A}_{(CEN)}$ for the proportional change in national median incomes by family size.

For income years 1989 and 1990, the estimator was:

$$\hat{\boldsymbol{A}}_{(MULT),t} = \hat{\lambda}_t \boldsymbol{A}_{(CEN)}$$

where $\hat{\lambda}_t$ is estimated by maximum-likelihood estimation. In other words, this estimator employs a single factor to inflate the growth in model uncertainty since the previous census, in place of the three parameters estimated by maximum-likelihood estimation. The results for income year 1989 were:

$$\hat{\boldsymbol{A}}_{(MULT),t} = \begin{pmatrix} 3.154 & 2.698 \\ 2.698 & 2.932 \end{pmatrix} \times 10^6$$

suggesting considerably greater growth in model error than accounted for by the assumptions underlying the original projection, which gave values:

$$\boldsymbol{A}_{(PROJ),t} = \begin{pmatrix} 1.000 & 0.827 \\ 0.827 & 0.870 \end{pmatrix} \times 10^6$$

In a manner analogous to the calculation for the 1980 census, the recently available 1990 census results give the following estimate:

$$\boldsymbol{A}_{(CEN)} = \begin{pmatrix} 1.631 & 1.444 \\ 1.444 & 1.494 \end{pmatrix} \times 10^6$$

This outcome is between the multiplicative results and the projection, although closer to the latter. In fact, some of the increase in the estimate of error based on the CPS sample estimates may be attributed to greater disagreement between the 1990 CPS and census state estimates of median income than expected based on CPS sampling variability, and further remarks on this point will be included in the final section. Note also how strikingly well these empirical results fit the multiplicative model. For example, choosing $\hat{\lambda}_t$ = 5.2955 gives:

$$\boldsymbol{A}_{(MULT)} = \begin{pmatrix} 1.631 & 1.394 \\ 1.394 & 1.514 \end{pmatrix} \times 10^6$$

4 EVALUATIONS

The availability of direct census estimates every 10 years affords a significant opportunity to evaluate and recalibrate the estimation technique. Fay, Nelson, and Litow (1993) describe a comparison of a previous version of the estimator to the 1980 census. Some of the features of the current model are an outgrowth of that earlier comparison.

Figures 1 - 5 compare the model to recently available estimates from the 1990 census. Overall, the results of the comparison are quite encouraging. For example, no estimate was in error by 10% or more, and only 7 were in error by 5% or more. These findings reflect only the first steps in a more complete analysis.

The next critical step, however, will be to react to a surprising finding reported in Section 3, namely, that the CPS sample estimates of the medians by state appear to differ from the CPS values by more than sampling error alone would suggest. This is in contrast to the previous comparison of the 1980 CPS and census, where sampling error alone appeared to account adequately for the observed differences. Consequently, some form of nonsampling error is possible, but a more systematic study of components of differences between the CPS and the census will be required to isolate the significant source or sources of these differences. The outcome of this investigation should provide a firmer basis to separate the issues of limitations of the model from possible nonsampling error in either the CPS or the census.

In turn, the results should permit assessment of a number of features of the current model:

1) The average error of the model predictions.
2) Whether errors are differential for certain classes of states, e.g., small vs. large, rapidly changing vs. static, lower income vs. higher, etc.
3) Whether errors cluster geographically.
4) Whether modification of the current predictors would yield significant improvement in prediction.

The census data permit assessment of the current model but also offer the occasion for consideration of more significant changes for subsequent years. A number of these are described in the next section.

In addition to relying on the census for evaluation, work on alternative models, such as the hierarchical Bayes model described by Datta, Fay, and Ghosh (1991), has addressed methods to obtain estimates of individual and aggregate measures of performance from the sample estimates when census data are not available. The 1990 census data should help to calibrate these procedures for future use. (Unfortunately, these procedures may be adversely affected by nonsampling error producing differences between the expected values of the CPS and the census medians at the state level, so that understanding sources of nonsampling error is a critical step here as well.)

5 FUTURE PLANS

Implemented a year at a time, the current model and its predecessor has produced a series spanning income years 1974 to 1991 without taking any advantage of the longitudinal or time series nature of this problem. Malay Ghosh and others are collaborating with the Census Bureau on improvements that attempt to address this aspect.

The current model relies simply on observed relationships that appear to be quite linear, without taking advantage of any specific knowledge about income distributions. Possibly, a more explicit parametric model for the income distribution may represent a fruitful alternative. On the other hand, the utility of such efforts would have to be balanced against requirements of parsimony imposed by the relatively small sizes of the CPS state samples.

As noted at the end of Section 4, another area of potential research is to attempt to improve measures of error from the model for the intercensal period. Recent research in fully Bayes procedures may prove promising for estimation of error.

[1] This paper reports results of research undertaken in part by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

Bailey, W., Hazen, L., and Zabronsky, D. (1993), "State Metropolitan Area, and County Income Estimation," in *Indirect Estimators in Federal Programs,* Federal Committee on Statistical Methodology, U. S. Office of Management and Budget, pp. (3-1)-(3-30).

Datta, G., Fay, R. E., and Ghosh, M. (1991), "Hierarchical and Empirical Multivariate Bayes Analysis in Small Area Estimation," in *Proceedings of the Annual Research Conference,* Washington, DC: U. S. Bureau of the Census, pp. 63-79.

Fay, R. E. (1986), "Multivariate Components of Variance Models as Empirical Bayes Procedures for Small Domain Estimation," in *Proceedings of the Survey Research Methods Section,* Washington, DC, American Statistical Association, pp. 99-107.

______ (1987), "Application of Multivariate Regression to Small Domain Estimation," in *Small Area Statistics, An International Symposium,* R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh, eds., New York: John Wiley & Sons, pp. 91-102.

Fay, R. E., Nelson, C. T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in *Indirect Estimators in Federal Programs,* Federal Committee on Statistical Methodology, U. S. Office of Management and Budget, pp. (9-1)-(9-17).

Fuller, W. A. and Harter, R. M. (1987), "The Multivariate Components of Variance Model for Small Domain Estimation," in *Small Area Statistics, An International Symposium,* R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh, eds., New York: John Wiley & Sons, pp. 103-123.

U.S. Bureau of the Census (1991), "Estimates of Median 4-Person Family Income by State: 1974-1989," Current Population Reports, Technical Paper No. 61, U.S. Government Printing Office, Washington, DC.