COMPARING ALTERNATIVE MODELS FOR USING DECENNIAL CENSUS DATA IN SAIPE STATE POVERTY ESTIMATES

Elizabeth T. Huang and William R. Bell
SRD, Bureau of the Census, Washington, DC 20233

KEY WORDS: **Small area estimation; Fay-Herriot model; Akaike's AIC**

1. Introduction

The Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program produces poverty estimates for various age groups for states, counties, and school districts. The state and county estimates are produced from various models applied to direct poverty estimates obtained from the March Supplement to the Current Population Survey (CPS). These models use predictor variables constructed from administrative data sources, demographic population estimates, and poverty estimates from the previous decennial census. The administrative data sources used include IRS tax file data, food stamp program participation data, and supplemental security income data from the Social Security Administration. The school district estimates are produced from simple synthetic updates to the previous census estimates with results controlled to the current county model-based estimates.

This paper focuses on use of decennial census poverty data in the state poverty models. SAIPE state production estimates were previously released for income years 1993 and 1995-98. The "income year," IY, refers to the year for which income is reported in the March CPS of the following year, the latter referred to as the "survey year." (IY 1994 was skipped due to technical difficulties with applying the models arising from the transition to the "new" CPS sample derived from a sample redesign based on 1990 census results.) The models that produced these previous SAIPE estimates used 1990 census results as the previous census data. The model-based estimates with documentation are available from the SAIPE web site at www.census.gov/hhes/www/saipe.html.

As of this writing, new SAIPE state estimates are in production for IY 1999. In developing these estimates we faced some interesting new issues regarding our use of census data because the Census 2000 long form poverty estimates are also for IY 1999. If the census estimates could be regarded as unbiased estimates of true poverty, then at the state level the census estimates would nearly provide truth for IY 1999, because sampling error in the census estimates is very small at the state level. Census poverty estimates are not truth, however, because the census is known to be subject to various nonsampling errors, which could here be thought of as biases. CPS poverty estimates have their own nonsampling errors, but these are believed to be less important than the nonsampling errors in the census estimates, due to the CPS using experienced interviewers to ask more detailed income questions than were asked of respondents on the census long form. There are also conceptual differences between what the CPS and census try to estimate as "poverty" (both for number of poor and poverty rates). Further discussion of CPS-census differences can be found on the SAIPE web site.

The presumed higher level of nonsampling error in the census estimates versus CPS, while the CPS state estimates contain significantly larger amounts of sampling error, makes it unclear how we can best use the 2000 census data in producing poverty estimates for IY 1999. One extreme would be to simply use the 2000 census estimates as the IY 1999 poverty estimates. This would reflect an assumption that the gain from greatly reducing sampling variance offsets the loss incurred from the census nonsampling errors. Another extreme would be to use the 2000 census data in our models in the same way the 1990 census estimates have been used in the models for previous income years. This requires minimal assumptions about census non-sampling error. As with the use of the 1990 census estimates in previous years, it merely reflects an assumption that whatever the 2000 census poverty estimates estimated is correlated with true poverty for the income year of interest (here 1999). The motivation for making no stronger assumption than this about the census estimates was clear when using the 1990 census estimates in a model for an income year later than 1989. But when the census and CPS estimates refer to the same year, the possibility that stronger assumptions about the census estimates could produce improved poverty estimates is worth considering.

This paper makes empirical comparisons of various approaches to using the census data in constructing CPS model-based poverty estimates for the census year (IY 1999 for the 2000 census). Because the CPS provides the official national level poverty estimates, we assume that the CPS provides unbiased estimates of true poverty. That is, we recognize sampling error in the CPS estimates but ignore their nonsampling errors. Under this assumption, we set up various alternative models that differ in how the census data are used. These alternative models include both the "census is truth" model and the SAIPE state production model as particular cases. We make statistical comparisons of these alternative models to determine which fits the CPS data best (allowing for the sampling error in the CPS estimates). We report statistical comparisons based on Akaike's AIC model selection criterion, though results from standard hypothesis tests (not reported) were similar. Other model selection criteria (e.g., BIC) could also be used.

Section 2 describes the SAIPE state model (Fay and Train 1997). Section 3 describes the alternative state poverty models compared here. Section 4 gives the empirical results of the model comparisons for census IYs 1989 and 1999 for poverty ratios for four age groups: 0-4, 5-17, 18-64, and 65 and over. As an extension we also present results for IYs1989-1993 and 1995-1998 for the four age groups. Section 5 compares prediction error variances between the current model and one of the alternative models for IYs 1989 and 1999. On the surface, the empirical results presented in Sections 4 and 5 seem favorable to using models with only the census estimates as a regression variable in the census years. Despite this, we concluded it would be more appropriate to continue to use the full model that also includes the regression variables drawn from administrative records data. Section 6 discusses the rationale for this conclusion and describes the Bayesian approach actually used for the IY 1999 SAIPE state estimates.

2. SAIPE State Model

Fay and Train (1997) developed models for CPS state poverty ratios for IY 1993. For 1995 and subsequent years, a model of the same form was used, with updated data and with some revisions to the predictor variables. The model applies to the CPS direct state estimates of poverty ratios (in percent), defined for each state and each age group being modeled as 100 times the ratio below:

$$\frac{\text{CPS estimated number in poverty in the age group}}{\text{CPS estimated population in the age group}}. \quad (1)$$

The ratio in (1) differs from a poverty rate because the denominator in (1) includes all persons in the age group, whereas for a poverty rate it would include only persons in the poverty universe. For discussion of the nature of this difference, which is generally small, see the SAIPE web site.

We model poverty ratios for age groups 0-4, 5-17, 18-64, and 65 and over. For age 5-17 the numerator of (1) is restricted to related children 5-17 in families in poverty. Estimates for related children 5-17 in families are important because they are used in Title 1 fund allocations by the U.S. Department of Education. SAIPE also produces estimates for total 5-17 children in poverty. The results for this group are quite similar to those for related children 5-17 and so are not discussed here.

The state model developed by Fay and Train (1997) for a given income year can be written as follows:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, 51 \qquad (2)$$

$$Y_i = \beta_0 + x_i'\beta + \gamma \mathrm{CenRes}_i + u_i \qquad (3)$$

$$u_i \sim \text{i.i.d. } N(0, \sigma_u^2)$$

$$e_i \sim \text{ind. } N(0, v_i) \qquad (4)$$

In (2) y_i denotes the CPS poverty ratio in (1) expressed in percent for state i (including DC as a "state" for estimation purposes), Y_i denotes the corresponding true poverty ratio, and e_i denotes the sampling error. Regression variables in the model (3) for Y_i include an intercept term, the vector x_i of current year predictors constructed from administrative records, and "census residuals" (explained shortly). β_0, β, and γ are the corresponding regression parameters. We assume that the sampling variances v_i are known though in practice they are estimated. Direct estimates of sampling error variances are produced by the successive difference replication method via VPLX (Fay and Train 1995). These estimates are then smoothed by a sampling error model (Otto and Bell 1995) to produce the $v_{i.}$

The "census residuals" in (3) derive from fitting the regression model analogous to (3), but without the $\gamma \mathrm{CenRes}_i$ term, to the census data. For IY 1999 this is

$$\mathrm{Cen}_i = \eta_0 + x_{i,99}'\eta + z_i \qquad (5)$$

$$z_i \sim \text{i.i.d } N(0, \sigma_z^2)$$

where in (5) Cen_i is the Census 2000 estimate for state i, and $x_{i,99}$ represents the same set of administrative records regression variables as in (3), but taking on values for the census IY 1999. There is no sampling error component in (5) because the sampling variance of the census long form state estimates is very low. Upon fitting (5) by ordinary least squares (OLS), the census residuals result from replacing η by its estimated value $\hat{\eta}$, i.e.,

$$\mathrm{CenRes}_i = \mathrm{Cen}_i - \hat{\eta}_0 - x_{i,99}'\hat{\eta}.$$

For the years prior to IY 1999, we derive the "census residuals" using 1990 census data and the regression variables $x_{i,89}$ that take on values for IY 1989.

Given the regression variables x_i, and the sampling variances v_i , model (3) can be estimated by Gaussian maximum likelihood (ML), or restricted ML (REML). It can also be given a Bayesian treatment (Bell 1999). In production, estimates of the v_i are iteratively updated with estimation of the regression parameters β and the model error variance σ_u^2 (Fay and Train 1997). We do this because the sampling error model we use allows for dependence of the sampling variances on the level of poverty. The true level of poverty is $Y_{i,}$, which is unobserved, so for this purpose it is estimated by the regression fit.

Given estimates of the model parameters (e.g., by ML or REML) model-based (empirical Bayes) estimates of the true poverty ratios are obtained from

$$\hat{Y}_i = h_i y_i + (1 - h_i)(\hat{\beta}_0 + x_i^{/}\hat{\beta} + \hat{\gamma}\text{CenRes}_i) \quad (6)$$

where $h_i = \sigma_u^2/[\sigma_u^2 + v_i]$, and β_0, β, γ, and σ_u^2 are estimated as noted above. These empirical Bayes estimates were used before IY 1997. Starting with IY 1997, the Bayesian treatment of the model was adopted using a flat prior for all model parameters. The Bayesian approach was adopted to more appropriately account for uncertainty about σ_u^2 (Bell 1999).

3. Alternative State Poverty Ratio Models

For each age group we consider various alternative models for Y_i in the context of the complete model for the CPS data given by (2)–(4). The sampling error model (effectively, the v_i) stays the same across the models being compared for a given age group. The idea is to use the CPS data for a given year to assess how best to use the census estimates for that year, with particular interest given to the census IYs 1989 and 1999. The set of models compared includes a model that assumes the census essentially provides the true poverty ratios, as well as the current production model which uses the "census residuals" and the administrative records variables. Some of the other models considered fall between these two models in regard to making different assumptions about the nature of measurement errors in the census estimates relative to the CPS estimates.

The first five models below (A1-A5) use only the census estimates (Cen) in the model for Y_i. Models A1–A3 and A5 allow for linear biases and/or random measurement errors (the u_i) in Cen_i as an estimate of Y_i. The other three models (B-D) use either census data or census residuals or both, as well as the administrative records regression variables x_i .

Model A1: $Y_i = \alpha_0 + \alpha_1\text{Cen}_i + u_i$

Model A2: $Y_i = \alpha_1\text{Cen}_i + u_i$

Model A3: $Y_i = \text{Cen}_i + u_i$

Model A4: $Y_i = \text{Cen}_i$ ("census is truth")

Model A5: $Y_i = \alpha_1\text{Cen}_i$

Model B: $Y_i = \beta_0 + x_i^{/}\beta + \alpha_1\text{Cen}_i + u_i$

Model C: The production model given by (3).

Model D:
$Y_i = \beta_0 + x_i^{/}\beta + \alpha_1\text{Cen}_i + \gamma\text{CenRes}_i + u_i$

Note that Model D is not estimable in the census years (IY 1989 or 1999) due to colinearity–CenRes_i is a linear function of Cen_i and x_i when all refer to the same year. Also, for this reason Models B and C provide equivalent fits in the census years.

4. Empirical Model Comparisons

We compared the fit of the alternative models described in Section 3 to the CPS state poverty ratio estimates for each age group and each income year for which we had data, paying particular attention to the results for the census IYs 1989 and 1999. In this section we present results obtained from models using the original, not updated sampling variances. Results using the updated sampling variances are provided in an internal report. The results are similar and lead to the same conclusions.

We compared model fits using Akaike's (1973) AIC, which is defined (for models estimated by ML) as

$$\text{AIC} = -2\times\text{Log-likelihood} + 2\times(\text{\# of parameters}).$$

The model with the minimum value of the AIC is preferred. We also performed Wald chi-squared tests of joint significance of the regression coefficients β in Model B as another way of checking whether the administrative records regressors were needed in the models. (Note: Testing the significance of β in Model C is not of interest because a model that uses only census residuals as a regressor is not plausible.) The chi-squared tests and AIC comparisons we performed led to essentially the same conclusions. Here we present just the results of the AIC comparisons.

Table 1 presents the AIC-preferred model for each age group for each income year considered: 1989-1993 and 1995-1998 (using 1990 census data), and 1999 (using 2000 census data). We notice from Table 1 that for the census years – IYs 1989 and 1999 – for all four age groups the preferred model is one of the A-models (which use only the census data and drop the x_i variables). For age 5-17 the "census is truth" model A4 is preferred for both census years. This result also holds for age 65+ though only for IY 1999. Also, for IY 1999 for age 0-4 the preferred model is A3, which says the census estimates are unbiased but contain random measurement error. For the other cases in the census years the preferred model involves a linear adjustment of the census estimates.

Table 1. AIC-Preferred State Poverty Ratio Models for IYs 1989-1993, 1995-1998, and 1999

a. Using 1990 census data

IY	age 0-4	age 5-17	age 18-64	age 65+
1989	A1	A4	A5	A5
1990	C	A1	C	A5
1991	C	D	C	A1
1992	C	C	B	A4
1993	C	D	C	A1
1995	C	C	C	C
1996	C	C	C	C
1997	D	C	D	C
1998	A3	C	D	C

b. Using 2000 census data

1999	A3	A4	A5	A4

For later IYs (1990-1998), for which the 1990 census estimates get progressively more and more out of date, we generally found the current production model C using both the census residuals and the administrative records variables to be the AIC-preferred model. Age 65 and over was an exception. For this age group the poverty ratio appears fairly stable over some years, so using either of the linear adjustment models A1 or A5 was usually preferred for 1990-93. The current production model C was favored for age 65+, however, for IYs 1995-1998. It is worth noting that the CPS data for these years was obtained under a new CPS design.

The interesting feature of the results from Table 1 is the general AIC preference for one of the A-models in the census years and the general preference for model C in the non-census years (with the exceptions noted for age 65+). Of course, these results do not imply that we believe the A-models are exactly true in the census years. In particular, the results (for age 5-17 in both census years and for age 65+ in IY 1999) do not imply that in these cases we really believe that the census estimates were exactly equal to the true poverty ratios. The AIC comparisons simply indicate a preferred choice of a prediction model among those models being compared. Put another way, the data simply provide no evidence in the census years that the models benefit from including the administrative records variables x_i. We return to this issue of implications of the results in Section 6.

5. Comparing State Prediction Error Variance Estimates from Two Models for the Census Years

In this section, we compare state prediction error variance estimates from Models A1 and B for the census IYs 1989 and 1999. Recall that, in census years, Model C has the same predictions and hence the same prediction error variances as Model B. Note that these comparisons of prediction error variances between Models A1 and B (or C) would not generally be meaningful for non-census years (except for age 65+), since the data show that the simpler Model A1 is inappropriate for those years.

To facilitate interpretation of these comparisons we make the calculations using the same value of the model error variance σ_u^2 for the two models, and similarly use the same sampling variances v_i. We do this to avoid having the prediction error variance comparisons distorted by differences in the estimates of σ_u^2 or in the updates of the v_i. It isolates the comparisons of the prediction error variances to comparisons of the part coming from error in estimating regression parameters (the second term in (7) below).

Given the AIC preference for the A-models in the census year (A1 being the most general such model), we use the estimated value of σ_u^2 from Model A1. We also use Model A1 to update the values of the sampling error variances v_i as discussed in Section 2 (using REML estimation of the model). These values of v_i and σ_u^2 are valid for use in Model B because the model comparison results of Section 4 suggest that in census years the additional regression variables x_i that distinguish Model B from Model A1 are unnecessary.

Let $\mathbf{x}_i$ be the full vector of the regression variables for any given model, including the intercept term, and, as appropriate, the regression variables from the administrative data x_i , and the census poverty rate or "census residuals." Let α be the corresponding full vector of regression parameters. For example, in the full Model B,

$$\mathbf{x_i}' = (1,\ x_i',\ \mathrm{Cen}_i) \text{ and } \alpha' = (\beta_0,\ \beta',\ \alpha_1\ ).$$

The variance of the prediction error for the poverty ratio Y_i for state i, given the model error variance σ_u^2, is as follows:

$$\mathrm{Var}(Y_i\,|y,\ \sigma_u^2) = \sigma_u^2(1-h_i) + (1-h_i)^2\,\mathbf{x_i}'\mathrm{Var}(\hat{\alpha})\mathbf{x_i} \quad (7)$$

where $h_i = \sigma_u^2 / (\sigma_u^2 + v_i)$, as noted earlier. We use the

Bayesian posterior mean of σ_u^2 from Model A1 in (7). For the census IY 1999 these posterior means are 7.29, 2.93, 0.43, and 0.71 for the age groups 0-4, 5-17, 18-64, and 65 and over, respectively. The expression (7) does not account for additional uncertainty due to estimating σ_u^2. Bell (1999) notes, however, that (*i*) in the Bayesian approach the contribution from this additional uncertainty is quite small in almost all cases, and (*ii*) in these cases the fully Bayesian results are very close to what results when the posterior mean of σ_u^2 is used in (7). (The contribution to uncertainty from estimating σ_u^2 is not small, though, when σ_u^2 is estimated by ML or REML.)

Given the values of the model error variances and the updated sampling error variances, we used (7) to compute the prediction error variances under Models A1 and B. We compared the prediction error variances by computing their percentage difference relative to the results from Model A1:

$$\% \text{ difference} = 100\times[V(Y_i)_B - V(Y_i)_{A1}] / V(Y_i)_{A1}.$$

Table 2 shows the frequency distributions of these relative percentage differences for IY 1999 for the four age groups. Similar results were obtained for IY 1989.

Table 2. Frequency distributions over states of the relative percentage difference of the prediction error variances from Models B and A1 for IY 1999

Percent	age 0-4	age 5-17	age 18-64	age 65+
0-5	20	12	6	8
5-10	11	12	16	9
10-20	8	12	12	12
20-30	5	8	6	4
30-40	5	3	5	6
40-50	0	2	3	2
50-60	0	0	0	2
60-70	0	0	1	2
70-80	0	0	0	0
80-90	0	1	0	0
90+	2	1	2	6

For IYs 1989 and 1999, the relative percentage differences are positive for all four age groups for all 51 states. Thus, Model A1, with fewer regression variables, yields lower prediction error variances because it involves estimating fewer regression parameters. The relative percentage differences for many states were small, but for others were not negligible, exceeding 20 percent for a number of states and exceeding 40 or 50 percent in some cases (with results varying over the age groups).

6. Postscript–Production of Estimates for IY 1999

The model comparisons of Section 4 were originally intended to provide a framework for deciding which models to use for the IY 1999 production estimates. Subsequent to obtaining the model comparison results for IY1989, but before getting the results for IY 1999, we rethought this framework. We then decided that it would be inappropriate to decide in favor of using one of the reduced (A) models for IY 1999 on the basis of model comparisons for that year. We instead thought it more appropriate to use the 2000 census data in our models for IY 1999 the same way we previously used the 1990 census data.

The thinking behind this decision was as follows. Generally speaking, with respect to fitting the full Model B and a reduced model such as Model A1 and comparing the results for the census IY (1999 or 1989), one of three situations will arise:

1. The regression coefficients on the x_i variables are statistically significant.

2. The regression coefficients on the x_i variables are not statistically significant, and in fact their estimates are close to zero and they have "relatively small" standard errors.

3. The regression coefficients on the x_i variables are not statistically significant but have "relatively large" standard errors (and thus their estimates may not be, practically speaking, very close to zero).

In situation 1 we clearly would choose to use the full Model B because the data provide evidence that the x_i variables are needed in the model for the census income year. In situation 2 the data provide evidence that the x_i variables are not needed in the model for the census income year, and so we have evidence in favor of the reduced model (say A1). However, since the standard errors of the coefficients on the x_i are small, the estimation of these coefficients will not contribute appreciably to the prediction error variances of the full model, and so the full and reduced models will produce very similar prediction results (point predictions and variances). In situation 3 the data do not provide strong evidence that the x_i variables are not needed, the data simply fail to provide evidence that these variables are needed. However, in this situation results from the full and reduced models, particularly prediction error variances, would be substantially different.

Note that it is only by estimating and doing the statistical comparisons of the full and reduced models that we could determine that we are not in situation 1. Having estimated the full model to determine this, i.e., that we have no evidence from the data that the x_i variables are needed, it seems statistically inappropriate to the use the reduced model and quote prediction error variances conditional on the reduced model being true. This ignores statistical uncertainty inherent in the estimation of the full model and the subsequent testing of it against the reduced model. This seems especially pertinent for situation 3 since there the data don't provide evidence in favor of the reduced model, they simply fail to provide evidence that the full model is needed. In situation 2 it may be more appropriate to proceed conditionally with the reduced model, since there we have evidence in its favor, but also in this situation it will not make much difference if we do this – there will be very little to gain from using the reduced model.

Subsequent to obtaining the results reported here, we obtained comments on an earlier version of this paper from Rod Little (University of Michigan) and Alan Zaslavsky (Harvard University). This included comments related to the "model choice conundrum" just discussed. The reviewers suggested shrinking estimates of the regression coefficients on the administrative records regression variables towards zero on the grounds that we expected them to be less relevant in the census IY 1999. We in fact implemented this general idea by using an informative prior on these coefficients with a prior mean of zero (while maintaining a flat prior on all the other coefficients). The prior covariance matrix of these regression coefficients was drawn from model fitting results for IY 1989, with an inflation factor (multiplying the standard deviations by 2) to reflect uncertainty about how well the results from IY 1989 would translate to 1999. The result was some reduction in the prediction error variances over straight use of the original full model B with a fully noninformative prior.

Disclaimer: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

References

Akaike, H. (1973), "Information Theory and an Extension of the Likelihood Principle," in the *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Czaki, Budapest: Akademia Kiado, 267-287.

Bell, William R. (1999), "Accounting for Uncertainty About Variances in Small Area Estimation," Bulletin of the International Statistical Institute, 52nd Session, Helsinki, 1999.

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 Government Statistics Sections of the American Statistical Association, pp 154-159.

Fay, R.E., and Train, G.F. (1997), "Small Domain Methodology for Estimating Income and Poverty Characteristics for States in 1993," Proceeding of the Government Statistics and Social Statistics Section, Alexandria, VA, American Statistical Association. p.183-188.

National Research Council (2000a), *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Method*, Report of the Panel on Estimates of Poverty for Small Geographic Areas, Constance F. Citro and Graham Kalton, editors,.Committee on National Statistics, Washington, D.C., National Academy Press.

National Research Council (2000b), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*, Report of the Panel on Estimates of Poverty for Small Geographic Areas, Constance F. Citro and Graham Kalton, editors. Committee on National Statistics. Washington, D.C.: National Academy Press.

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.