MULTIVARIATE COMPONENTS OF VARIANCE MODELS AS EMPIRICAL BAYES PROCEDURES FOR SMALL DOMAIN ESTIMATION

Robert E. Fay, U.S. Bureau of the Census

1. INTRODUCTION

Frequently, sample surveys produce estimates of characteristics for the total population or for large domains within the population, but interest extends beyond this level of detail to small areas or domains. Often, design-based estimators for the small domains are unsatisfactory because of small sample sizes and associated high sampling errors. Most attempts to address this problem have involved the use of auxiliary information outside the survey itself. Some, such as Sarndal (1984), have proposed estimators incorporating the auxiliary information entirely within a design-based framework. Most developments in this area, however, have relied upon combining the sample survey information with analytic models for the population. A review article of Purcell and Kish (1979) summarizes many of these modeling approaches.

Work of Ericksen (1973, 1974) on the use of linear regression for small domain estimation represents one example of modeling and has formed the basis for a number of applications. In his original description of this method, a sample survey is assumed to provide unbiased sample estimates, Y_i's, for many or all of the small domains of interest. The Y_i's are presumed to have estimable, even though generally large, sampling errors. Auxiliary information, X_{ij}, is presumed to be available for each domain i and for a specific number of characteristics j. In this formulation, the X_{ij}'s are presumed to be measured without error, but they are also assumed to be typically imperfect linear predictors of $E(Y_i)$, the underlying characteristic measured by Y_i. The predicted values from the linear regression of the Y_i's on the X_{ij}'s provide small domain estimates for the population. Ericksen also presented an estimator of the mean-squared error of prediction for these small domain estimates.

Fay and Herriot (1979) described some modifications to the regression approach related to the methods presented in this paper. By drawing parallels to the James-Stein estimator when the Y_i's have equal sampling variances, they proposed an empirical Bayes estimator for the situation of unequal sampling variances. In the case of domains with no sample data, the estimator is the regression estimate, as before. For each domain with a sample estimate, Y_i, however, the empirical Bayes estimator takes the form of a composite, that is, a weighted combination of Y_i with the predicted value from the regression, where the weights for domain i depend on both the sampling variance of Y_i and the overall fit of the regression to the sample estimates.

This paper considers extensions of this approach to the instance in which some of the independent variables, the X_{ij}'s, are themselves stochastic. For example, some of the independent variables may have also been measured by the same or a different sample survey. Under limited circumstances, use of the previous estimators developed for fixed X_{ij}'s still proves appropriate, but a more general solution requires that the effect of random variation in the X_{ij}'s be explicitly included in the model. Specifically, the method described here involves treating all stochastic variables, both the Y_i's and any stochastic X_{ij}'s, as dependent variables in a multivariate regression expressed as a multivariate components of variance model. Separate variance components are included in the model to represent sampling variability in the survey estimates and to represent the lack of fit or model error of the model. By estimating the unknown variance components and then applying the best linear unbiased estimator (Harville 1976, 1977) as if the estimated variance components were known, an empirical Bayes estimator results. This estimator again is in the form of composite between the sample estimates and the regression estimates, although the composite is now multivariate.

A potential application of this idea is to situations in which small domain estimates are of interest for more than one characteristic. In many instances, a more effective estimate of each characteristic emerges than if each had been treated as a separate estimation problem. In this case, the multivariate formulation appears a natural extension of the regression method.

When only one characteristic is of interest, the multivariate formulation is less obvious; yet, this approach generally yields more effective estimates than does the univariate regression of the Y_i's on the non-stochastic X_{ij}'s. The gain comes from use, under the model, of the information contained in the stochastic X_{ij}'s.

An earlier paper (Fay 1985) outlined this methodology, but did not contain the empirical results reported here. The methods are related to and preceded by work of Fuller and Harter (1985). The major difference between their approach and the one presented here reflects the level of analysis. Fuller and Harter modeled the data at the level of the individual observation, and derived small domain estimates as a consequence, whereas the models in this paper apply directly at the level of the small domain. Although the approach in the Fuller and Harter paper may be more effective in situations satisfying their conditions, the models in this paper permit more general sample designs. The two approaches are approximately equally well suited to applications in which the auxiliary information is available only at the level of the small domain or cannot be matched at the individual level to the sample survey data. Additionally, models at the level of the small domain are applicable to statistics, such as medians, defined at the level of the small domain and not fitting into the formulation at the individual level. Further remarks on the relationship between the models considered in this paper and those of Fuller and Harter appear at the end of the paper.

The next section of this paper provides the background for the application to be presented in this paper, the estimation of median income for 4-person families by state. Section 3 describes the formulation of a components of variance model for the problem. Section 4 presents the empirical results and discusses alternative approaches to estimating the components of variance by combining information over years. The concluding section suggests other ways in which these and similar models may be applied to problems of small domain estimation.

2. THE PROBLEM: ESTIMATION OF MEDIAN INCOME FOR 4-PERSON FAMILIES BY STATE

2.1 Basic Objectives and Sources of Data

The Social Security Administration, Department of Health and Human Services, administers a program, Low Income Home Energy Assistance, to aid low income families with their energy costs. Eligibility for the program varies by state and is determined by a formula employing an estimate of the median income of 4-person families by state (and the District of Columbia) for the most recently available year.

The Current Population Survey (CPS) provides direct sample survey estimates of median income for 4-person families annually at the national level. The CPS data may also be tabulated to produce separate estimates by state. At the state level, however, the effect of sampling variability is pronounced: in a few states the coefficient of variation (c.v., that is, standard error divided by expected value) is only 2 or 3 percent, but in most states it falls in the range of 6 to 10 percent.

The models to be described for this problem have employed two auxiliary sources of information: results from the decennial censuses of population, and estimates from the Bureau of Economic Analysis (BEA). The census provides income values for the year immediately preceding the census year. At the state level, the census estimates of median income for 4-person families have negligible sampling error. Thus, the census values of median income for 1979 constitute an obvious choice for inclusion in a model of current state medians.

BEA produces annual estimates of per capita income for states and other geographic areas on the basis of aggregate statistics on earnings, transfer payments, etc. The estimates furnish an overall indication of relative income among states and especially change over time for individual states, even though per capita income is only indirectly related to median income of 4-person families. The BEA estimates, although subject to revision and conceptual differences from census income, are essentially free from sampling error.

Up to three independent variables, X_{i1}, X_{i2}, and X_{i3}, are employed in this paper as predictors of $E(Y_i)$, where Y_i represents the expected value of median income from the CPS. X_{i1} takes the value 1, serving as the independent variable corresponding to the constant term;

$$X_{i2} = (BEA(t)_i/BEA(c)_i)\ Cen(c)_i \tag{2.1}$$

where $BEA(t)_i$ and $BEA(c)_i$ denote the values of BEA per capita income in state i for the current income year and census income year, respectively, and $Cen(c)_i$ represents the median income for the year measured by the census; and $X_{i3} = Cen(c)_i$ denotes the unadjusted census median. X_{i2} represents an attempt to adjust the census values of median income for change since the census according to the proportional growth of BEA per capita income.

2.2 The Earlier Model

A few years before the 1980 census, a regression estimator was developed for this problem. The CPS sample estimates, Y_i, for the median income of 4-person families by state were fitted by a linear regression with X_{i1} and X_{i2}, where the latter were 1969 median incomes adjusted by change in BEA per capita income. A composite estimate was formed from the sample estimates Y_i and the fitted regression estimates. The selection of the independent variables and the assessment of this model were based on fitting 1970 census values for 1969 median income using X_{i2} expressed as the 1960 census values adjusted for change from 1959 to 1969 in BEA income, as in (2.1). The observed average error in fitting the 1970 census values was used as an estimate of the model error for subsequent years, instead of estimating the model error from the sample data, as suggested by Fay and Herriot (1979).

If the composite estimator combining the sample and regression estimates fell beyond one standard error of the sample estimate, the final estimate was constrained to lie within one standard error of the sample value. This procedure was originally included to limit the individual risk to any of the states in case the model failed seriously in a few isolated instances. The application described by Fay and Herriot (1979) also incorporated this constraint. A drawback of the procedure here, however, was that the constraint increased the year-to-year variability in the estimates.

Comparison of the estimates from the model for income year 1979 with the 1980 census afforded an opportunity to assess the merits of the estimator. Several conclusions emerged from this review:

i. Although the original model attempted to distinguish among states according to population size in assessing the average model error, the distinction proved ineffective for the 1979 estimates. The average error across all states, however, agreed well with the previous experience derived from fitting the the 1970 census. Instead, it now appeared appropriate to presume a single level of model error across all states.

ii. The addition of X_{i3}, the unadjusted median from the previous census, improved the predictive power of the regression. The form of X_{i2} in (2.1) proportionally adjusts the census median for changes in BEA income; inclusion of X_{i3} permits this adjustment to be softened in case of any "regression toward the mean," which might arise if BEA income did not perfectly measure the proportional changes in the income distribution.

iii. Use of the constraint of one standard deviation from the sample estimate did not reduce the number of extreme errors compared to the same model without the constraint. Had the model produced extreme errors in a few states, the constraint would have been effective, but, in the absence of extreme errors, the constraint failed to be beneficial. This experience suggested removing the constraint.

These three changes in the model could have been expected to yield improvements in the performance for post-1980 predictions. Nonetheless, the remaining year-to-year variability provided the impetus to attempt additional modifications. The CPS sample estimates for 3- and 5-person families represented an additional source of information. Strong statistical relationships among the median incomes of 3-, 4-, and 5-person families at the state level appear in both the 1970 and 1980 censuses. In 1980, the ratio of medians for the 4-person to 3-person families ranged only from 1.072 to 1.162 at the state level. This ratio fell between 1.089 and 1.135 for two-thirds of the states. Furthermore, some of the differences among states appeared structural, since almost all of the higher ratios appeared in states with relatively large Black populations. Similarly, the association between 4- and 5-person medians at the state level was almost as high.

Consequently, if CPS obtained 3- and 5-person medians without sampling error, they would have made excellent predictors to add to the regression. The sampling error of these medians is comparable to those for 4-person families, however. The next section describes a model incorporating the 3- and 5-person medians while accounting for the effect of sampling error.

3. THE MODEL FOR FAMILY MEDIANS BY STATE

As stated in the first section, the formulation of the components of variance model includes stochastic predictors as dependent variables in the regression. In a preceding paper (Fay 1985), the original approach envisioned 4-, 3-, and 5-person family medians each as separate dependent variables. When this model was applied, however, it became evident that the greater sampling variability of the 5-person medians and especially the instability of the estimated sampling error for these medians were problematic. Instead, the information from the 5-person medians was incorporated into the model by averaging the 5- with the 3-person median at the state level, with weights 1/4 and 3/4, respectively. The weights were chosen on the basis of the approximate 1 to 3 ratio for the respective sample sizes. Thus, the regression was reduced to two dependent variables per state. Although the notation presented here is specific to the problem at hand, the methods are applicable to more than two variables per domain.

Let $\underset{\sim}{\mu}(i) = (\mu(i)1, \mu(i)2)'$ denote the true values of the medians in state i, where $\mu(i)1$ is the true median for 4-person families and $\mu(i)2$ denotes the weighted average computed as 3/4 the 3-person median and 1/4 the 5-person median. Assume that $\mu(i)k$, k=1,2, has been sampled from a superpopulation with expected value $\underset{\sim}{X}(i)(k)\ \underset{\sim}{\beta}(k)$. For k=1, the elements $X(i)(1)j$ of $\underset{\sim}{X}(i)(1)$ are X_{ij}, j = 1, 2, 3, from the previous tion. For k=2, the elements of $\underset{\sim}{X}(i)(2)$ are formed similarly, with $Cen(c)_i$ in (2.1) and as $X(i)(2)3$ equal to the weighted average of the census 3- and 5-person family medians. Assume that the distributions of $\underset{\sim}{\mu}(i)$ about their expected values are $N(\underset{\sim}{0}, \underset{\sim}{A})$ and independent over i.

The corresponding CPS estimates $\underset{\sim}{Y}(i)$ for state i are assumed $N(\underset{\sim}{\mu}(i), \underset{\sim}{D}(i))$, with known sampling covariance matrix $\underset{\sim}{D}(i)$. The sampling errors will also be assumed independent across states.

Under this superpopulation model, the unconditional distribution of the Y's may be expressed as a mixed components of variance model. Setting

$$\underset{\sim}{Y} = (Y(1)1, Y(1)2, Y(2)1, \ldots)',$$

a column vector of 102 elements (for 50 states and DC), and $\underset{\sim}{X}$ to be the 102 by 6 matrix whose first rows are given by

$$\underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X}(1)(1) & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{X}(1)(2) \\ \underset{\sim}{X}(2)(1) & \underset{\sim}{0} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix},$$

the model is

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{b} + \underset{\sim}{e} \tag{3.1}$$

where $\underset{\sim}{b}$ represents random effects arising from the presumed sampling of the $\underset{\sim}{\mu}(i)$ from the superpopulation (i.e., the bias terms of the regression model when the true state medians are considered fixed), and $\underset{\sim}{e}$ denotes sampling errors. The covariance matrix of $\underset{\sim}{b}$ is assumed to be in the block diagonal form

$$\underset{\sim}{A}^* = \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{0} & \underset{\sim}{0} & \cdot & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{A} & \underset{\sim}{0} & \cdot & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{A} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \end{bmatrix},$$

and that of $\underset{\sim}{e}$ in the form

$$\underset{\sim}{D}^* = \begin{bmatrix} \underset{\sim}{D}(1) & \underset{\sim}{0} & \underset{\sim}{0} & \cdot & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{D}(2) & \underset{\sim}{0} & \cdot & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{D}(3) & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \end{bmatrix}.$$

If $\underset{\sim}{A}$ were known, the BLUE estimate of the fixed effects would be given by

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}\underset{\sim}{Y}. \tag{3.2}$$

Consequently, the BLUE of the superpopulation means would be given by $\underset{\sim}{X}\hat{\underset{\sim}{\beta}}$. The objective, however, is to estimate the actual true medians,

$$\underset{\sim}{\mu} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{b},$$

which is a linear combination of the fixed and random effects. Again, assuming $\underset{\sim}{A}^*$ were known, the BLUE of $\underset{\sim}{\mu}$ is given by

$$\hat{\underset{\sim}{\mu}} = \underset{\sim}{X}\hat{\underset{\sim}{\beta}} + \underset{\sim}{A}^*(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}(\underset{\sim}{Y}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}}) \tag{3.3}$$

(Harville 1976). Equation (3.3) gives the BLUE as the regression estimate $\underset{\sim}{X}\hat{\underset{\sim}{\beta}}$ plus the multivariate residual $(\underset{\sim}{Y}-\underset{\sim}{X}\hat{\underset{\sim}{\beta}})$ times a "shrinkage factor," actually a matrix, $\underset{\sim}{A}^*(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}$. Harville (1977) noted that this estimator corresponds to the posterior mean of an analogous Bayes model with a uniform prior on the fixed effects, $\underset{\sim}{\beta}$. Because of the block-diagonal form of $\underset{\sim}{A}^*$ and $\underset{\sim}{D}^*$ in this specific model, the residuals of both components in state i will be incorporated in the estimation of each component in (3.3), unless $\underset{\sim}{D}_{(i)}$ is a multiple of $\underset{\sim}{A}$, but the residuals of each state do not affect the estimation in other states in (3.3). Indeed, it is through the form of (3.3) that the information represented in the CPS estimates of the 3- and 5-person family medians can can improve the estimation of the 4-person family medians.

In fact, $\underset{\sim}{A}$ is generally not known. Substitution of a value of $\underset{\sim}{A}$ estimated from the data into (3.2) and (3.3) converts (3.3) into a parametric empirical Bayes estimator, in the sense of Morris (1983).

The next section examines the relative merits of estimating $\underset{\sim}{A}$ directly from the data or from past experience, in this case from the fit of the model to census data. In other applications, however, direct estimation may be the only available choice. Direct estimation from the data is based on maximum likelihood estimation in the next section, although alternatives such as MINQUE may also be considered in other applications.

The log-likelihood, ignoring additive constants not depending upon the parameters, is

$$L(\underset{\sim}{A},\underset{\sim}{\beta};\underset{\sim}{Y}) = -(1/2)\log(\det(\underset{\sim}{A}^*+\underset{\sim}{D}^*)) - (1/2)(\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}(\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{\beta}). \quad (3.4)$$

One method to obtain the maximum likelihood estimate is to choose test values of $\underset{\sim}{A}$, compute $\hat{\underset{\sim}{\beta}}$ from (3.2) and then examine the resulting log-likelihood from (3.4). The maximum can be obtained by search procedures, which are particularly easy for this problem because of the small dimension of $\underset{\sim}{A}$. Harville (1977) reviewed more systematic approaches to finding the maximum.

4. IMPLEMENTATION OF THE MODEL

4.1 Comparisons to 1980 Census

Since the 1980 census provides state medians essentially free from sampling error, evaluation of the fit of the regression to the census values gives a relatively precise determination, $\underset{\sim}{A}_0$, of $\underset{\sim}{A}$. Here, of course, 1970 census medians are used as the census values in the definition of the independent variables, $\underset{\sim}{X}_{(i)(k)}$. The elements of $\underset{\sim}{A}_0$ are 3.08×10^5 for A_{011}, the variance of the model error term for 4-person family medians, 2.86×10^5 for A_{022}, the corresponding variance for the weighted combination of 3- and 5-person medians, and 2.63×10^5 for A_{012}. As expected, $\underset{\sim}{A}_0$ gives a high correlation between the two components, .89.

The diagonal elements of $\underset{\sim}{D}_{(i)}$, the sampling variances for the census medians, were estimated according to a methodology documented in an internal memorandum (Fay and Burkhead 1985). Because the 4 person medians and the weighted average of 3- and 5-person medians were derived from mutually exclusive sets of households, their sampling covariance was taken to be zero. It is possible that a correlation in fact exists from clustering effects in the CPS design, but this correlation is certainly small and decidely much less than that of $\underset{\sim}{A}$.

Maximum likelihood estimates, computed directly from the 1980 CPS data and using the estimated $\underset{\sim}{D}_{(i)}$ as if they were known, give A_{11} as 1.15×10^5, A_{22} as 3.26×10^5, and A_{12} as 1.93×10^5. The estimated correlation is 1.00. Although the MLE estimate, $\hat{\underset{\sim}{A}}$, is somewhat different from $\underset{\sim}{A}_0$, the difference in the the log-likelihood (3.4), multiplied by two, is only 1.13, indicating no statistically significant difference. As will be shown later, the likelihood surface is quite flat for $\underset{\sim}{A}$ for this problem, implying wide confidence regions for $\hat{\underset{\sim}{A}}$.

The shrinkage factor, $\underset{\sim}{A}^*(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}$, in (3.3) makes considerable use of the sample estimates for the weighted 3- and 5-person family medians in estimating the 4-person family medians, because both $\underset{\sim}{A}_0$ and $\hat{\underset{\sim}{A}}$ have such different correlations from $\underset{\sim}{D}_{(i)}$. Table 1 compares the 1980 census values for the median income of 4-person families in 1979 with three sets of estimates: estimates derived under the earlier regression methodology using only CPS data for 4-person medians, (3.3) based upon $\underset{\sim}{A}_0$, and (3.3) based upon $\hat{\underset{\sim}{A}}$. The second set of estimates is dependent through $\underset{\sim}{A}_0$ on results from the 1980 census, while the first and third were derived entirely independently of the 1980 census.

Generally, the two sets of new estimates are closer to each other than to the previous methodology. Table 1 shows substantial improvements over the earlier approach. One of the first set of estimates is in error by over 10.0 percent and 11 addition estimates are in error by 5.0 percent or more. The new estimates have no errors over 10.0 percent and only three each in the range of 5.0 to 9.9 percent.

4.2 Estimates for Income Years 1981-1984

Table 1 shows $\hat{\underset{\sim}{A}}$ to do quite well against $\underset{\sim}{A}_0$. Even though the two estimated covariance matrices are quite dissimilar, their difference has little effect on the log-likelihood function. The flatness of the log-likelihood function implies the possibility of large variation in $\hat{\underset{\sim}{A}}$ from year to year. A more stable estimate of $\underset{\sim}{A}$, although not necessarily superior, may be derived by multiplying $\underset{\sim}{A}_0$ by the square of the proportional growth in national median income, thus assuming approximately the same average relative error of the regression model over time. Table 2 presents estimates based on these projected error components; table 3 gives those based on the MLE. The projected error components give a somewhat smoother series of estimates, although the contrast between the sets are not dramatic.

Table 4 compares the projected components and MLE estimates for these four years. The MLE estiestimates vary substantially over time, but the overall test of difference from the projected components based on a chi-square test with three degrees of freedom is not significant in each case. Again, the likelihood surface is quite flat. In 1984, the MLE of the correlation, r, drops substantially, but not necessarily significantly. Because of this low estimated r, the estimates in table 3 for 1984 based on the MLE of A make little use of the CPS estimates for the 3- and 5-person medians in predicting the 4-person medians. Use of the projected error components assures that the 3- and 5-person medians will be used to approximately the same degree in each year.

The estimates in table 2 based on projected components have been selected as the preferred set. Estimates from the 1986 CPS for income year 1985 will provide additional perspective on this choice, when they become available.

5. CONCLUDING REMARKS

The representation of the estimation problem in terms of a components of variance model allows auxiliary information with random error to be suitably combined into the estimation. The use of this information depends upon the form of the factor $\underset{\sim}{A}^*(\underset{\sim}{D}^*+\underset{\sim}{A}^*)^{-1}$ in (3.3). For small domain applications in which $\underset{\sim}{A}^*$ and $\underset{\sim}{D}^*$ take the same block diagonal form as in this example, the auxiliary information will have effect on the estimation of the k-th characteristic when the k-th row of $\underset{\sim}{A}(\underset{\sim}{D}_i+\underset{\sim}{A})^{-1}$ has nondiagonal elements. The importance of the auxiliary information in this instance resulted from the high correlation of the characteristics in the population and from a negligible correlation of the sample estimates about the true values. It is precisely this combination of circumstances that would favor application of this methodology to other problems.

Research is underway to adapt these methods to the estimation of median fair-market rent for 2-bedroom units for larger SMSA's and regions from the Annual Housing Survey (AHS). These medians are computed from a subset of rental units meeting a number of criteria. The situation differs somewhat from the CPS application since the 1980 census did not collect sufficiently detailed data to determine which units satisfy the definition of this quantity exactly. The census does provide related data, however. Additionally, rentals for 2-bedroom units not satisfying the definition and rentals for units of other sizes should be statistically related to the variable of interest at the SMSA level. The sample estimate for such related variables should have low sampling covariances with each other. This situation appears to offer the same potential gains from the components of variance approach as in the application to median incomes for 4-person families from CPS.

As mentioned in the introduction of this paper, Fuller and Harter (1985) proposed a similar use of components of variance models for small domain estimation, where the model is stated at the individual level. In some applications, particularly when the interest is in domain means and when auxiliary information can be matched at the individual level, their approach should be more effective than the one presented here. In cases in which modeling at the individual level is inappropriate or proves difficult, such as the estimation of medians, modeling at the level of the small domain becomes a feasible and effective alternative.

Acknowledgements. The author is grateful to Charles H. Alexander and Beverley D. Causey for thoughful comments and to Alice W. Coburn for assistance with typing.

References

Ericksen, E.P. (1973), "A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography 10, 137-160.

____ (1974), "A Regression Method for Estimating Population Changes for Local Areas," Journal of the American Statistical Association 69, 867-875.

Fay, R.E. (1985), "Application of Multivariate Regression to Small Domain Estimation," paper presented at the International Symposium on Small Area Statistics, Ottawa, Ontario, Canada, May 22-24.

____ and Burkhead, D. (1985), "Summary of Research Completed on Estimating Median Income for 4-Person Families by State," undated draft memorandum to Roger A. Herriot, Chief, Population Division, U.S. Bureau of the Census.

____ and Herriot, R. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association 74, 269-277.

Fuller, W.A. and Harter, R.M. (1985), "The Multivariate Components of Variance Model for Small Domain Estimation," paper presented at the International Symposium on Small Area Statistics, Ottawa, Ontario, Canada, May 22-24.

Harville, D.A. (1976), "Extension of the Gauss-Markov Thereom to Include the Estimation of Random Effects," Annals of Statistics 4, 384-395.

____ (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," Journal of the American Statistical Association 72, 320-338.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," Journal of the American Statistical Association 78, 47-55.

Purcell, N.J. and Kish, L. (1979), "Estimation for Small Domains," Biometrics 35, 365-384.

Sarndal, C.E. (1984), "Design-Based Versus Model Dependent Estimation for Small Domains," Journal of the American Statistical Association 79, 624-631.

Table 1. Comparison of 1979 Estimates of Median Income of 4-Person Families

State	1980 Census	Original Method	Comp of Var Methods A_0	Comp of Var Methods $\hat{A}$	Percent Error Orig.	Percent Error A_0	Percent Error $\hat{A}$
ME	18319	18074	18810	18842	-1.3	2.7	2.9
NH	22027	22335	22626	22559	1.4	2.7	2.4
VT	19424	19314	20297	20426	-.6	4.5	5.2
MA	23772	23786	23809	23714	.1	.2	-.2
RI	22107	21636	22022	21908	-2.1	-.4	-.9
CT	25712	24410	25623	25585	-5.1	-.3	-.5
NY	22669	21082	22313	22438	-7.0	-1.6	-1.0
NJ	26014	24640	25440	25337	-5.3	-2.2	-2.6
PA	22266	22314	22657	22384	.2	1.8	.5
OH	23279	22528	22986	23051	-3.2	-1.3	-1.0
IN	23014	22614	22489	22499	-1.7	-2.3	-2.2
IL	25410	24265	24611	24602	-4.5	-3.1	-3.2
MI	25111	24422	24716	24687	-2.7	-1.6	-1.7
WI	23320	23518	24108	24124	.8	3.4	3.4
MN	24044	24409	23840	23884	1.5	-.8	-.7
IA	22351	22567	22561	22451	1.0	.9	.4
MO	21891	21294	21688	21753	-2.7	-.9	-.6
ND	20511	19520	19910	20031	-4.8	-2.9	-2.3
SD	18674	19209	19346	19370	2.9	3.6	3.7
NE	21438	20749	20922	20807	-3.2	-2.4	-2.9
KS	22127	22848	22404	22409	3.3	1.3	1.3
DE	23627	21184	22007	21940	-10.3	-0.9	-7.1
MD	26203	24686	25306	25152	-5.8	3.4	-4.0
DC	21862	21310	22184	22202	-2.5	1.5	1.6
VA	22757	22976	23076	22895	1.0	1.4	.6
WV	20214	18876	19210	19308	-6.6	-5.0	-4.5
NC	19772	19648	19507	19447	-.6	-1.3	-1.6
SC	19944	20154	19498	19532	1.1	-2.2	-2.1
GA	20668	21578	20902	20692	4.4	1.1	.1
FL	21086	20757	20798	21086	-1.6	-1.4	.0
KY	19685	19138	19423	19422	-2.8	-1.3	-1.3
TN	19693	19437	19155	19143	-1.3	-2.7	-2.8
AL	19926	18613	19317	19532	-6.6	-3.1	-2.0
MS	18150	17672	18024	17948	-2.6	-.7	-1.1
AR	17893	18493	18149	18101	3.4	1.4	1.2
LA	21412	20166	20154	20229	-5.8	-5.9	-5.5
OK	20659	20852	20655	20770	.9	.0	.5
TX	22521	23416	22519	22429	4.0	.0	-.4
MT	20776	20051	20106	20068	-3.5	-3.2	-3.4
ID	19961	20429	20517	20407	2.3	2.8	2.2
WY	24641	22673	24236	24334	-8.0	-1.6	-1.2
CO	23757	25228	24370	24145	6.2	2.6	1.6
NM	19257	21032	19960	20007	9.2	3.7	3.9
AZ	21924	23000	22733	22731	4.9	3.7	3.7
UT	21572	21250	21234	21280	-1.5	-1.6	-1.4
NV	24438	25457	24351	24278	4.2	-.4	-.7
WA	24394	24410	24341	24234	.1	-.2	-.7
OR	22688	24031	23308	23242	5.9	2.7	2.4
CA	24752	25109	24393	24570	1.4	-1.5	-.7
AK	31018	31037	30012	30020	.1	-3.2	-3.2
HI	24966	24582	25115	25091	-1.5	.6	.5

Table 2 Estimates of Median Income of 4-Person Families, Using Projected Census Components of Variance

State	1979 (Census)	1981 Est.	1982 Est.	1983 Est.	1984 Est.	Percent Change 79-84	79-81	81-82	82-83	83-84
ME	18319	21433	22842	23852	26237	43.2	17.0	6.6	4.4	10.0
NH	22027	25980	26881	29337	33255	51.0	17.9	3.5	9.1	13.4
VT	19424	23080	24053	24466	26645	37.2	18.8	4.2	1.7	8.9
MA	23772	28409	29441	33258	36731	54.5	19.5	3.6	13.0	10.4
RI	22107	25655	27116	29093	32066	45.0	16.0	5.7	7.3	10.2
CT	25712	30431	32077	35474	39070	52.0	18.4	5.4	10.6	10.1
NY	22669	26224	27615	30140	32665	44.1	15.7	5.3	9.1	8.4
NJ	26014	30498	31547	35141	39096	50.3	17.2	3.4	11.4	11.3
PA	22266	25607	26691	28221	29573	32.8	15.0	4.2	5.7	4.8
OH	23279	26413	27771	28556	30779	32.2	13.5	5.1	2.8	7.8
IN	23014	25567	26827	27545	30302	31.7	11.1	4.9	2.7	10.0
IL	25410	29493	30736	31615	33126	30.4	16.1	4.2	2.9	4.8
MI	25111	28862	29362	30230	32365	28.9	14.9	1.7	3.0	7.1
WI	23320	27349	28372	28846	30622	31.3	17.3	3.7	1.7	6.2
MN	24044	27864	28582	30652	33817	40.6	15.9	2.6	7.2	10.3
IA	22351	25824	26780	26766	28650	28.2	15.5	3.7	-.1	7.0
MO	21891	25276	26173	27602	30050	37.3	15.5	3.5	5.5	8.9
ND	20511	24443	25557	27012	28901	40.9	19.2	4.6	5.7	7.0
SD	18674	21326	22505	22849	25391	36.0	14.2	5.5	1.5	11.1
NE	21438	24995	26080	26253	28752	34.1	16.6	4.3	.7	9.5
KS	22127	25353	26480	27769	30330	37.1	14.6	4.4	4.9	9.2
DE	23627	27730	29078	31320	33809	43.1	17.4	4.9	7.7	7.9
MD	26203	30909	31843	35223	38132	45.5	18.0	3.0	10.6	8.3
DC	21862	25059	26217	28949	31104	42.3	14.6	4.6	10.4	7.4
VA	22757	27052	28111	30591	33480	47.1	18.9	3.9	8.8	9.4
WV	20214	22730	23764	23265	25316	25.2	12.4	4.5	-2.1	8.8
NC	19772	23227	24257	25944	27995	41.6	17.5	4.4	7.0	7.9
SC	19944	22578	23258	26719	27810	39.4	13.2	3.0	6.3	12.5
GA	20668	24470	25719	26874	29623	43.3	18.4	5.1	4.5	10.2
FL	21086	24410	25023	26800	28858	36.9	15.8	2.5	7.1	7.7
KY	19685	23011	24361	24245	25815	31.1	16.9	5.9	-.5	6.5
TN	19693	22915	24039	24313	26603	35.1	16.4	4.9	1.1	9.4
AL	19926	22773	24309	25014	26595	33.5	14.3	6.7	2.9	6.3
MS	18150	21020	22102	22135	23660	30.4	15.8	5.1	.1	6.9
AR	17893	20672	21503	21822	23075	29.0	15.5	4.0	1.5	5.7
LA	21412	25108	25904	27554	28430	32.8	17.3	3.2	6.4	3.2
OK	20659	24712	25700	26809	28856	39.7	19.6	4.0	4.3	7.6
TX	22521	26574	27761	28884	31031	37.8	18.0	4.5	4.0	7.4
MT	20776	24512	24980	25465	26072	25.5	18.0	1.9	1.9	2.4
ID	19961	23159	24111	26244	25499	27.7	16.0	4.1	.6	5.2
WY	24641	29174	29435	29256	29752	20.7	18.4	.9	-.6	1.7
CO	23757	28310	29264	31526	34154	43.8	19.2	3.4	7.7	8.3
NM	19257	22355	23093	23974	25468	32.3	16.1	6.0	1.2	6.2
AZ	21924	25494	26924	27586	29431	34.2	16.3	5.6	2.5	6.7
UT	21572	24700	25981	25528	27497	27.5	14.5	5.2	-1.7	7.7
NV	24438	28321	29160	30488	31059	27.1	15.9	3.0	4.6	1.9
WA	24394	28091	29228	30102	31585	29.5	15.2	4.0	3.0	4.9
OR	22688	25832	27356	27497	28633	26.2	13.9	5.9	.5	4.1
CA	24752	20502	29603	31734	33711	36.2	15.2	3.9	7.2	6.2
AK	31018	36958	37234	42867	44017	41.9	19.2	.7	15.1	2.7
HI	24966	24324	30236	32030	33445	34.0	17.5	3.1	5.9	4.4

Table 3 Estimates of Median Income of 4-Person Families, Using Estimated Components of Variance from CPS

State	1979 (Census)	1981 Est.	1982 Est.	1983 Est.	1984 Est.	Percent Change 79-84	79-81	81-82	82-83	83-84
ME	18319	21386	22980	24058	26055	42.2	16.7	7.5	4.7	8.3
NH	22027	25908	26623	29040	33701	53.0	17.6	2.8	9.1	16.1
VT	19424	22936	24400	24448	26770	37.8	18.1	6.4	.2	9.5
MA	23772	28226	29624	33664	36652	54.2	18.7	5.0	13.6	8.9
RI	22107	25743	27037	29024	31832	44.0	16.4	5.0	7.3	9.7
CT	25712	30345	32490	35613	39573	53.9	18.0	7.1	9.6	11.1
NY	22669	26349	27541	30152	33214	46.5	16.2	4.5	9.5	10.2
NJ	26014	30460	31501	35044	39588	52.2	17.1	3.4	11.2	13.0
PA	22266	25628	26636	28199	29815	33.9	15.1	3.9	5.9	5.7
OH	23279	26460	27713	28442	30583	31.4	13.7	4.7	2.6	7.5
IN	23014	25862	26613	27432	30751	33.6	12.4	2.9	3.1	12.1
IL	25410	29413	30923	31786	31902	25.5	15.8	5.1	2.8	.4
MI	25111	28774	29242	30172	31781	26.6	14.6	1.6	3.2	5.3
WI	23320	27145	28572	28751	30774	32.0	16.4	5.3	.6	7.0
MN	24044	27780	28352	30686	33655	40.0	15.5	2.1	8.2	9.7
IA	22351	25850	26780	26946	28530	27.6	15.7	3.6	.6	5.9
MO	21891	25269	25983	27601	29953	36.8	15.4	2.8	6.2	8.5
ND	20511	24497	25698	26913	28682	39.8	19.4	4.9	4.7	6.6
SD	18674	21392	22535	22877	25419	36.1	14.6	5.3	1.5	11.1
NE	21438	24984	26094	26005	28657	33.7	16.5	4.4	-.3	10.2
KS	22127	25430	26562	27801	30713	38.8	14.9	4.5	4.7	10.5
DE	23627	27636	29250	31476	34048	44.1	17.0	5.8	7.6	8.2
MD	26203	30828	32048	35445	38664	47.6	17.7	4.0	10.6	9.1
DC	21862	25220	25764	28876	31093	42.2	15.4	2.2	12.1	7.7
VA	22757	26878	28313	30779	33146	45.7	18.1	5.3	8.7	7.7
WV	20214	22858	23437	22625	25102	24.2	13.1	2.5	-3.5	10.9
NC	19772	23077	24325	26375	28322	43.2	16.7	5.4	8.4	7.4
SC	19944	22778	22806	24435	28051	40.6	14.2	.1	7.1	14.8
GA	20668	24259	25953	26849	28820	39.4	17.4	7.0	3.5	7.3
FL	21086	24552	24761	26709	28188	33.7	16.4	.9	7.9	5.5
KY	19685	22914	24672	24330	25220	28.1	16.4	7.7	-1.4	3.7
TN	19693	22848	24042	24289	26121	32.6	16.0	5.2	1.0	7.5
AL	19926	22873	24488	25129	27060	35.8	14.8	7.1	2.6	7.7
MS	18150	20935	21858	22158	23429	29.1	15.3	4.4	1.4	5.7
AR	17893	20619	21409	21872	22505	25.8	15.2	3.8	2.2	2.9
LA	21412	25155	25730	27553	29108	35.9	17.5	2.3	7.1	5.6
OK	20659	24599	25766	27111	30118	45.8	19.1	4.7	5.2	11.1
TX	22521	26632	27887	28729	32014	42.2	18.3	4.7	3.0	11.4
MT	20776	24378	24914	25373	26626	28.2	17.3	2.2	1.8	4.9
ID	19961	23118	24174	24329	25718	28.8	15.8	4.6	.6	5.7
WY	24641	29049	29485	29383	31432	27.6	17.9	1.5	-.3	7.0
CO	23757	28206	29467	31739	34117	43.6	18.7	4.5	7.7	7.5
NM	19257	22348	23850	24074	25046	30.1	16.1	6.7	.9	4.0
AZ	21924	25497	27079	27601	29067	32.6	16.3	6.2	1.9	5.3
UT	21572	24686	26101	25282	26898	24.7	14.4	5.7	-3.1	6.4
NV	24438	28235	29317	31015	31145	27.4	15.5	3.8	5.8	.4
WA	24394	28093	29132	30060	32167	31.9	15.2	3.7	3.2	7.0
OR	22688	25825	27668	27819	28847	27.1	13.8	7.1	.5	3.7
CA	24752	28612	29511	31738	33325	34.6	15.6	3.1	7.5	5.0
AK	31018	36985	36820	42068	43760	41.1	19.2	-.4	14.3	4.0
HI	24966	29193	30440	32338	34000	36.2	16.9	4.3	6.2	5.1

Table 4 Comparison of Projected Components and MLE's by Year
(All variance components shown divided by 10^5)

Year	4-Person	Wtd. Avg. 3- and 5-	Covariance	r	Difference 2*Log-Lklhd
1981					
Projected	4.154	3.747	3.500	.89	.49
MLE	2.069	2.591	2.315	1.00	
1982					
Projected	4.590	4.003	3.803	.89	
MLE	7.082	12.038	9.234	1.00	3.19
1983					
Projected	5.125	4.412	4.219	.89	
MLE	8.774	9.749	9.248	1.00	1.67
1984					
Projected	5.819	5.244	4.900	.89	5.57
MLE	18.058	13.616	2.160	.14	