소지역통계 추정

2002. 11. 28.

통계기획국 조사관리과

SMALL AREA ESTIMATION

Dal Ho Kim

DEPARTMENT OF STATISTICS

KYUNGPOOK NATIONAL UNIVERSITY

OUTLINE

Part I : Introduction to Small Area Estimation

- Definition and Historical Background
- Examples
- Synthetic and Composite Estimation (Without Auxiliary Information)
- Synthetic and Composite Estimation (With Auxiliary Information)
- Bayesian Interpretation of Composite Estimation
- Hierarchical and Empirical Bayes Methods

Part II : Some Data Analysis

- Estimation of Median Income of Four-Person Families

Part I : Introduction to SAE

Definition

Small Area (or Local Area) :
A small geographical area such as a county, municipality or a census division.

Could be a "small domain", i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area.

Early History

11th Century England
17th Century Canada

Based on census or administrative records aiming at complete enumeration.

Sample Survey Data :

Provide reliable estimators of totals and means for large areas or domains.

Direct survey estimators for a small area, based on data only from the sample units in the area, yield large SE's due to unduly small size of the sample in the area.

The reason behind this is that the original survey was designed to provide specific accuracy at a much higher level of aggregation than that for small areas.

This makes it a necessity to "borrow strength" from related areas through implicit or explicit models that connect the small areas to find more accurate estimates for a given area, or simultaneously, for several areas.

Demand for small area statistics : From both public and private sectors.

Examples

I. SAIPE(http://www.census.gov/hhes/www/saipe)

US Federal Govt : Need for small area income and poverty estimates

Bill H.R. 1645 passed by the House of Representatives on Nov 21, 1993.

This bill requires the Secretary of Commerce to produce and publish at least every two years beginning in 1996, current data related to the incidence of poverty in the United States.

Specifically, the legislation states that "to the extent feasible", the Secretary shall produce estimates of poverty for states, counties and local jurisdictions of government and school districts. For school districts, estimates are to be made of the number of poor children aged 5-17 years. It also specifies production of state and county estimates of the number of poor persons aged 65 or over.

Currently the decennial census is the only source of income distribution and poverty data for households, families and persons for such small geographic areas.

These statistics for small areas are used by a broad range of customers including policy makers at the state and local levels as well as the private sector for allocation of federal and state funds, federal funds being more than $30 billion for the fiscal year 1994.

Use of 1990 census data pertaining to the economic situation in 1989 is questionable as it does not adequately reflect the current situation.

II. Estimation of Per Capita Income(PCI) for Several Small Places

(Ref. Fay and Herriot (JASA, 1979, 269-277))
Census Bureau provides the Treasure Dept. with PCI estimates and other statistics for state and local govts. receiving funds under the General Revenue Sharing Program.

Treasure Dept. uses these statistics to determine allocations to local govts. within the diff. states by dividing the corresponding state allocations.

Earlier Approach (of the 70's)

Current PCI Estimate

$$= \text{PCI Estimate of 1969 (Based on 1970 census)} \times \frac{\text{Current Administrative PCI estimate}}{\text{1969 administrative PCI estimate}}$$

Problem : Among 39,000 local govt. units, 15,000 were for places having fewer than 500 persons in 1970. CV's based on these sample estimates ranged from 13 to 30 percent.

Fay and Herriot : Empirical Bayes (EB) estimates.

Weighted average of the census sample estimates and a "synthetic" estimate. (Regression Estimate)

Fit a linear regression to the sample estimates of PCI using as independent variables "county average", "tax return data for 1960" and "data on housing". (Fay-Herriot model)

III. Prediction of Areas under Corn and Soybeans for 12 Counties in North-Central Iowa

(Ref. Battese, Harter and Fuller (JASA, 1988))
Based on farm-interview data as well as LANDSAT satellite data.

Areas of corn and soybeans in 37 sample segments (each segment was about 250 hectares) of 12 counties were determined by interviewing farm operators. Based on LANDSAT readings obtained auxiliary data which classify the crop cover for all pixels (a term for "picture element" about 0.45 hectares) in the 12 counties. BHF use a "nested error regression model" involving random small area effects and the segment-level data.

IV. Estimates of Median Income of Four-Person Families for the 50 States and the District of Columbia

The U.S. Dept. of Health and Human Services(HHS) has a direct need for such data at the state level in formulating its energy assistance program for low-income families.

The basic source of data is the annual demographic supplement to the March sample of the Current Population Survey(CPS) which provides the median income of four-person families for the proceeding year.

Direct use of CPS estimates is usually undesirable because of large CV's associated with them.

The Bureau of the Census started using a regression method - $\log y$ since the latter part of the 70's which uses in addition to the current CPS medians, the corresponding medians from the recentmost decennial census.

It was replaced later by a more sophisticated empirical Bayes methodology of Fay (1987).

V. Estimation of Cancer Mortality Rates

(Clayton & Kalder, Biometrics, 1987; Tsutakana, Biometrics, 1985)

Mortality due to lung cancer : Annual frequency in a small or average-sized city is quite low and information from a single city is very limited.

Use information from several cities with differing mortality rates that will yield better estimates of true mortality rates than the raw rates based on individual cities.

Example : Lung cancer mortality for males aged 45 -64 in Missouri cities.

Let Y_i = # of deaths due to lung cancer in the city in 1972-81.

Model : (i) $Y_i \sim$ Poisson$(n_i p_i)$, n_i =size of ith city;

(ii) $\theta_i = \log \frac{p_i}{1-p_i}$ iid $N(\mu, \sigma^2)$

(iii) $\mu \sim$ uniform$(-\infty, \infty)$; $\sigma^2 \sim$ inverse gamma.

Synthetic and Composite Estimation (without Auxiliary Information)

y_{ij} =characteristic of interest for the jth unit in the ith local area $(j = 1, \cdots, N_i, i = 1, \cdots, m)$

Let $y_{ij} (j = 1, \cdots, n_i)$ denote the characteristics corresponding to the n_i sampled units.

Population Mean :

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \quad (i = 1, \cdots, m)$$

Direct Estimate (Sample Mean) :

$$\bar{y}_{is} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (i = 1, \cdots, m)$$

Synthetic Estimates :

$$\bar{y}_s = \sum_{i=1}^{m} n_i \bar{y}_{is} / \sum_{i=1}^{m} n_i$$

Composite Estimates :

$$w_i \, \bar{y}_{is} + (1 - w_i) \bar{y}_s$$

Q. How to choose the weights w_i?

Both design- and model-based approaches have been proposed.

Often $w_i = n_i / N_i$ or $\sum_{i=1}^{m} n_i / \sum_{i=1}^{m} N_i$.

A Model-Based Justification of the Estimator

$$\frac{n_i}{N_i}\,\bar{y}_{is} + (1 - \frac{n_i}{N_i})\,\bar{y}_s$$

Let y_{ij} $(j = 1, \cdots, N_i;\ i = 1, \cdots, m)$ be iid with mean θ and variance 1. Then

$$E(\bar{Y}_i | y(s)) = \frac{n_i}{N_i}\,\bar{y}_{is} + \frac{N_i - n_i}{N_i}\theta.$$

But the BLUE of θ is $\bar{y}_s$. Hence, the BLUP of $\bar{Y}_i$ is

$$\frac{n_i}{N_i}\,\bar{y}_{is} + \left(1 - \frac{n_i}{N_i}\right)\bar{y}_s.$$

Later, we shall motivate this estimator from a Bayesian point of view.

Synthetic and Composite Estimation (with Auxiliary Information

y_{ij} =characteristic of interest for the jth unit in the ith local area $(j = 1, \cdots, N_i;\ i = 1, \cdots, m)$

$\boldsymbol{x}_{ij}$ =vector of auxiliary characteristics for the jth unit in ith local area $(j = 1, \cdots, N_i;\ i = 1, \cdots, m)$

For simplicity, consider scalar x_{ij}'s. Often $x_{ij} = x_i,\ \forall j$.

Population Mean :

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \quad (i = 1, \cdots, m)$$

Direct Estimator (Ratio Estimator) : $\left(\bar{y}_{is}/\bar{x}_{is}\right) \bar{X}_i$
$(\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij};\ \bar{x}_{is} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij})$

Synthetic Estimator : $\hat{\bar{Y}}_i^{RS} = \left(\bar{y}_s/\bar{x}_s\right) \bar{X}_i$

Composite Estimator : $\frac{n_i}{N_i}\bar{y}_{is} + \left(1 - \frac{n_i}{N_i}\right) \frac{\bar{y}_s}{\bar{x}_s} \bar{X}_i'$
$(\bar{X}_i' = (N_i - n_i)^{-1} \sum_{j=n_i+1}^{N_i} x_{ij})$

A model based justification of

$$\frac{n_i}{N_i}\,\bar{y}_{is} + \left(1 - \frac{n_i}{N_i}\right)\frac{\bar{y}_s}{\bar{x}_s}\bar{X}_i'$$

Consider the model

$$y_{ij} \overset{ind}{\sim} (b\,x_{ij},\ \ \sigma^2\,x_{ij})$$

BLUE of b is obtained by minimizing

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - b\,x_{ij})^2/(\sigma^2\,x_{ij}) \text{ wrt } b$$

i.e. $\quad b^2\sum_i\sum_j x_{ij} - 2b\sum_i\sum_j y_{ij} + \sum_i\sum_j x_{ij}^2$ wrt b

$$\hat{b} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}} = \frac{\bar{y}_s}{\bar{x}_s}$$

$$E\left[\bar{Y}_i | y_{i1}, \cdots, y_{in_i},\ i = 1, \cdots, m\right] = \frac{n_i}{N_i}\,\bar{y}_{is} + b\,\frac{N_i - n_i}{N_i}\bar{X}_i'$$

Substitute $\hat{b}$ for b.

(Ref : Holt, Smith and Tomberlin, JASA, 1979, 405-410)

Bayesian Interpretation of Composite Estimators

First consider the case when there is no auxiliary information.

As before, let $y_{i1}, \cdots, y_{in_i} (i = 1, \cdots, m) | \theta \overset{iid}{\sim} N(\theta, 1)$

Suppose also $\theta \sim$ uniform$(-\infty, \infty)$.

As before, let $y_{i1}, \cdots, y_{in_i}$ $(i = 1, \cdots, m)$ denote characteristic of interest corresponding to the sampled units. Then

$$
\begin{aligned}
& f(\theta \mid y_{i1}, \cdots, y_{in_i} \ (i = 1, \cdots, m)) \\
& \quad \propto f(y_{i1}, \cdots, y_{in_i} \ (i = 1, \cdots, m), \theta) \\
& \quad \propto \exp\left[-\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \theta)^2 \right] \\
& \quad \propto \exp\left[-\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_s)^2 + n_T (\bar{y}_s - \theta)^2 \right]
\end{aligned}
$$

where $n_T = \sum_1^m n_i$.

Hence $\theta \mid y(s) \sim N(\bar{y}_s, n_T^{-1})$

This implies

$y_{i\,(n_i+1)}, \cdots, y_{i\,N_i}\ (i = 1, \cdots, m) \,|\, y(s)$

is multivariate normal with common mean $E(\theta \,|\, y(s)) = \bar{y}_s$, common variance $V(\theta \,|\, y(s)) + 1 = n_T^{-1} + 1$, and pairwise common covariance $V(\theta \,|\, y(s)) = n_T^{-1}$

Consequently,

$N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \Big| y(s)$ is normal

with mean

$$E\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \Big| y(s)\right] = N_i^{-1}\left[n_i \bar{y}_{is} + (N_i - n_i)\bar{y}_s\right]$$

and

$$V\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \Big| y(s)\right] = \frac{N_i - n_i}{N_i^2} + \left(\frac{N_i - n_i}{N_i}\right)^2 n_T^{-1}.$$

In contrast, using the model based approach (which is also an empirical Bayes approach)

$V\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \Big| y(s)\right]$ is

estimated by $N_i^{-2}(N_i - n_i)$,

i.e. the uncertainty in estimating θ is not taken into account.

Auxiliary Information Case

Consider the regression model

(i) $y_{ij}|b \overset{ind}{\sim} N(b\,x_{ij}, \sigma^2\,x_{ij})$

where $x_{ij}(>0)$ and σ^2 are known.

(ii) $b \sim$ uniform$(-\infty, \infty)$.

Then $f(b|y(s)) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-b\,x_{ij})^2/x_{ij}\right]$.

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[(y_{ij}-b\,x_{ij})^2/x_{ij}\right]$$
$$= b^2\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij} - 2b\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{ij}^2/x_{ij}.$$

Hence $b|y(s) \sim N\left(\frac{\bar{y}_s}{\bar{x}_s}, \frac{\sigma^2}{n_T\,\bar{x}_s}\right)$.

Hence the joint posterior distribution of

y_{ij} $(j = n_i+1, \cdots, N_i\,;\ i = 1, \cdots, m)$ given $y(s)$

is multivariate normal with

$E(y_{ij}|y(s)) = E(b|y(s))x_{ij} = \frac{\bar{y}_s}{\bar{x}_s}x_{ij}$

$V(y_{ij}|y(s)) = V(b|y(s))x_{ij}^2 + \sigma^2 x_{ij}$

and $Cov(y_{ij}, y_{i'j'}|y(s)) = V(b|y(s))x_{ij}x_{i'j'}$

$(1 \le i \ne i' \le m,\ n_i+1 \le j \le N_i,\ n_{i'}+1 \le j' \le N_{i'})$

This implies

$$E\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}\Big|y(s)\right]$$
$$= N_i^{-1}\left[n_i\,\bar{y}_{is} + \frac{\bar{y}_s}{\bar{x}_s}\sum_{j=n_i+1}^{N_i} x_{ij}\right]$$
$$= N_i^{-1}\left[n_i\,\bar{y}_{is} + (N_i - n_i)\frac{\bar{y}_s}{\bar{x}_s}\bar{X}_{i'}\right]$$

(same as the one of Holt et. al (1979))
However,

$$V\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}\Big|y(s)\right]$$
$$= V\left[N_i^{-1}\sum_{j=n_i+1}^{N_i} y_{ij}\Big|y(s)\right]$$
$$= V(l|y(s))N_i^{-2}(N_i - n_i)^2\bar{X}_{i'}^2 + \sigma^2(N_i - n_i)^2\bar{X}_{i'}$$

in contrast to the model-based(or EB) estimator $\sigma^2(N_i - n_i)^2\bar{X}_{i'}$.

Hierarchical and Empirical Bayes Methods for SAE

Bayesian methods have been used quite extensively in recent years for solving small-area estimation problems. Particularly effective in this regard have been the hierarchical Bayes (HB) and empirical Bayes (EB) approaches, which are especially suitable for a systematic connection of local areas through models to "borrow strength" in a sensible way.

Suppose there are m local areas labelled $1, \cdots, m$.

Parameters of interest : $\theta_1, \cdots, \theta_m$
(e.g. local area totals or means)
Direct survey estimators : $\widehat{\theta}_1, \cdots \widehat{\theta}_m$

In addition, only area specific auxiliary data $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$ $(i = 1, \cdots, m)$ are available.

Bayesian Regression Model

I. $\widehat{\theta}_i | \theta_i \overset{ind}{\sim} N(\theta_i, V_i)$
II. $\theta_i \overset{ind}{\sim} N(\boldsymbol{x}_i^T \boldsymbol{b}, \tau^2)$

For now, V_i are assumed known.

Posterior:

$\theta_i | \widehat{\theta}_i \overset{ind}{\sim} N((1 - B_i)\widehat{\theta}_i + B_i \boldsymbol{x}_i^T \boldsymbol{b},\ V_i(1 - B_i))$
where $B_i = \frac{V_i}{V_i + tau^2}$.

Example (Fay and Herriot, 1979)

$\exp(\theta_i)$ = True PCI for the ith local area
$\exp(\hat{\theta}_i)$ = Direct estimate of PCI for the ith local area (= A, say)

Note: $A = B \times \frac{C}{D}$

B = PCI estimate of the ith local area from the recentmost decennial census
C = Current administartive estimate of PCI
D = Similar administrative estimate of PCI for the year preceding the recentmost decennial census

$x_{i1} = 1 \ \forall i$
$x_{i2} = log(\text{PCI})$ for the county where the ith local area belongs from the recentmost decennial census

This is the simplest version of the regression model considered by Fay and Herrior. More complex models include the value of owner-occupied housing from the recentmost decennial census, and the average adjusted gross income per exemption from the IRS returns in the year preceding the recentmost decennial census.

V_i (known) : Based on census returns.

Empirical Bayes Analysis

Estimate $\boldsymbol{b}$, τ^2 from the marginal distribution of $\hat{\theta}_i$'s.
Marginally, $\hat{\theta}_i \overset{ind}{\sim} N(\boldsymbol{x}_i^T\boldsymbol{b}, \tau^2 + V_i)$

Several procedures for estimating $\boldsymbol{b}$ and τ^2 :
(i) Fay and Herriot (1979) (ii) Morris (1983)
(iii) Prasad and Rao (1990) (iv) Cressie (1992)

Write
$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1.\cdots, \hat{\theta}_m)^T$
$\boldsymbol{X}^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$
$\boldsymbol{D} = \text{Diag}(V_1 + \tau^2, \cdots, V_m + \tau^2)$
Assume rank$(\boldsymbol{X}) = p < m$.

Writing $\hat{B}_i = \frac{V_i}{V_i + \hat{\tau}^2}$, EB estimator of $\boldsymbol{\theta} = (\theta_1.\cdots, \theta_m)^T$ is given by

$$\hat{\boldsymbol{\theta}}^{EB} = (\hat{\theta}_1^{EB}, \cdots, \hat{\theta}_m^{EB})^T,$$

where $\hat{\theta}_i^{EB} = (1 - \hat{B}_i)\hat{\theta}_i + \hat{B}_i\boldsymbol{x}_i^T\tilde{\boldsymbol{b}}(\hat{\tau}^2)$,
where $\tilde{\boldsymbol{b}}(\tau^2) = (\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}$.

Such an estimator can also be interpreted as an empirical best linear unbiased predictor (EBLUP).

Remark. Prasad and Rao (1990) find an approximation of the Bayes risk (they call it mean squared error) of $\hat{\boldsymbol{\theta}}^{EB}$ as an estimator of $\boldsymbol{\theta}$ unser squared error loss and the subjective prior $\theta_i \overset{ind}{\sim} N(\boldsymbol{x}_i^T\boldsymbol{b}, \tau^2)$.

Hierarchical Bayes Analysis

Hierarchical Model :
I. $\hat{\theta}_i|\theta_i \overset{ind}{\sim} N(\theta_i, V_i)$
II. $\theta_i \overset{ind}{\sim} N(\boldsymbol{x}_i^T\boldsymbol{b}, \tau^2)$
III. $\boldsymbol{b}$ and τ^2 are marginally independent with $\boldsymbol{b} \sim$ uniform(R^k) and $g(\tau^2) \propto 1$

Goal :
Find the posterior distribution of $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\theta}}$. Also compute $E(\theta_i|\hat{\boldsymbol{\theta}})$ and $V(\theta_i|\hat{\boldsymbol{\theta}})$.

In these days, it is much easier to carry out the HB analysis using the Gibbs sampler. As in Gelfand and Smith (1991), we use Rao-Blackwellized estimates for posterior means and variances.

Remark. It turns out that the EB and HB methods can quite often lead to comparable results especially in the context of point estimation. However, when it comes to the question of measuring standard errors associated with these estimates, the HB method has a clear edge over a naive EB procedure. The reason behind this is that a naive EB procedure fails to incorporates any uncertainty involved in estimating the unknown prior parameters, and thus leads to underestimates of the standard errors. In contrast, the HB method usually reports the posterior S.D.'s as errors associated with posterior means.

Part II : Some Data Analyses

Estimation of Median Income of Four-Person Families

The U.S. Dept of Health and Human Services provides energy assistance to low-income families. Eligibility for the program is determined by a formula where the most important variable is an estimate of the current median income of four-person families by states (the 50 states and the DC).

The Bureau of the Census, by an informal agreement, has provided such estimates to the HHS through a linear regression methodology since the later part of the 1970's.

Pre-1985 Approach

Sample estimates of the state medians for the most current year c, as well as the associated SE's are first obtained through Current Population Survey (CPS). These estimates are used as dependent variables in a linear regression procedure with the single predictor variable

Adjusted Census Median(c)
= [BEA PCI(c)/BEA PCI(b)] $\times$ Census Median(b)

along with an intercept term.

In the above, census median (b) represents the median income of four-person families in the state in the base year (the year preceding the census year) b from the most recently available decennial census. Also BEA PCI (c) and BEA PCI (b) represent respectively estimates of per-capita income produced by the Bureau of Economic Analysis of the U.S. Dept. of Commerce for the current year c, and the base-tear b respectively. Thus, adjusted census median (c) attempts to adjust the base-year census median by the proportional growth in the BEA PCI to arrive at the current year adjusted median.

In developing the model, the states were divided on the basis of the population into 4 groups of 12 or 13 states each. Model variances for each were computed as the average of squared residuals of 1969 census medians fitted by a linear regression with an intercept term with the adjusted census median for 1969 (with 1959 as the base-year) as a covariate.

Next a weighted average of the CPS sample estimate of the current median income and the regression estimate are obtained, weighting the regression estimate inversely proportional to the model variance, and the sample (CPS) estimate inversely proportional to the sampling variance.

Following the suggestion of Fay (1987), we use the census median for the base year (b) as a second independent variable. The idea is to adjust for any possible overstatement of the effect of change in BEA income in the median income of four-person families. Also, we did not divide the states into 4 groups, but lumped them together for "borrowing strength".

Let
θ_i = true median income of four-person families in 1979 for the ith state.
$\widehat{\theta}_i$ = CPS estimate of median income of four-person families in 1979 for the ith state.

Consider the hierarchical model given in p.22. Here
x_{i1} = adjusted census median income for the ith state;
x_{i2} = base year census median income for the ith state.

Now consider the HB estimates, EB estimates, CPS estimates, and the Bureau of the Census estimates, and compare them all against the 1979 census estimates, the "gold standard" under the following four criteria recommended by the panel on small area estimates of population and income set up by the committee on National Statistics.

Suppose $e_{i,TR}$ denotes the true (census) median income for the ith state in 1979 and e_i any estimate of $e_{i,TR}$ $(i = 1, \cdots, 51)$.

Then, for the estimate $e = (e_1, \cdots, e_51)^T$ of $e_{TR} = (e_{1,TR}, \cdots, e_{51,TR})^T$,

•Average Relative Bias $= (51)^{-1} \sum_{i=1}^{51} |e_i - e_{i,TR}|/e_{i,TR}$

•Average Squared Relative Bias $= (51)^{-1} \sum_{i=1}^{51} (e_i - e_{i,TR})^2/e_{i,TR}^2$

•Average Absolute Bias $= (51)^{-1} \sum_{i=1}^{51} |e_i - e_{i,TR}|$

•Average Squared Deviation $= (51)^{-1} \sum_{i=1}^{51} (e_i - e_{i,TR})^2$

The next two tables provide the conclusions.

A Multivariate Model

In addition to a change in the regression model used previously by the Bureau of the Census, Fay suggested also incorporation of median income of five- and three-person families as well because of the correlation in the median income of three-, four- and five-persion families

Fay really suggested a bivariate procedure where one uses (i) median income of fourperson families and (ii) $\frac{3}{4}$(median income of three-person families) $+\frac{1}{4}$(median income of five-person families).

We have discussed previously univariate EB and HB models, and have anlized the 1979 data. Now we proceed to analyze the data using multivariate models. We consider three bivariate models which include
(i) three- and four-person families
(ii) three- andfive-person families
(iii) Fay's models

In addition, we consider also the trivariate model.

For brevity, we describe fully only the trivariate model.

Let Y_{i1}, Y_{i2} and Y_{i3} denote respectively the sample (CPS) median incomes of four-, three- and five-person families. The corresponding true median incomes are denoted by θ_{i1}, θ_{i2} and θ_{i3} respectively. Let

$\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, $(i = 1, \cdots, 51)$;

$\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \theta_{i3})^T$, $(i = 1, \cdots, 51)$.

Also, let

$$\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_{51})^T; \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{51})^T.$$

Other notations:

x_{i11} = adjusted census median income for four-person families in the ith state;

x_{i21} = adjusted census median income for three-person families in the ith state;

x_{i31} = adjusted census median income for five-person families in the ith state;

x_{i12} = base year census median income for four-person families in the ith state;

x_{i22} = base year census median income for three-person families in the ith state;

x_{i12} = base year census median income for five-person families in the ith state.

Let

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & x_{i11} & x_{i12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{i21} & x_{i22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_{i31} & x_{i32} \end{bmatrix}$$

$i = 1, \cdots, 51.$

For bivariate models, we appropriately modify $\boldsymbol{Y}_i$, $\boldsymbol{\theta}_i$, $\boldsymbol{X}_i$. (Note in the bivariate case, $\boldsymbol{X}_i$ is 2×6).

Hierarchical Models:

I. $\boldsymbol{Y}_i|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{a} \overset{ind}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{V}_i)$

II. $\boldsymbol{\theta}_i|\boldsymbol{b}, \boldsymbol{a} \overset{ind}{\sim} N(\boldsymbol{X}_i\boldsymbol{b}, \boldsymbol{a})$

III. Marginally $\boldsymbol{b}$ and $\boldsymbol{a}$ are independent with $\boldsymbol{b} \sim$ uniform(R^p) and $\pi(\boldsymbol{a}) \propto 1$.

In the multivariate cases, we carry out the HB analysis using Gibbs sampler.

A Time Series Approach

Now we add a new feature to this problem. Suppose our target is to estimate the median income of four-person families in the year 1989 (The year 1989 is picked because it is then possible to compare the estimates with the census figures).

Based on the earlier approach, we will take 1979 as the base year, and then use 1989 CPS and auxiliary information from 1979 and 1989 to arrive at the estimates for 1989 leaving all the data from the intermediate years unused.

However, we have data from the intermediate years as well which suggest use of a time series modeling.

Basic Data : $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})^T$

Y_{ij1} = sample (CPS) median incomes of four-person families for state i in year j
Y_{ij2} = sample (CPS) median incomes of three-person families for state i in year j
Y_{ij3} = sample (CPS) median incomes of five-person families for state i in year j

True Mean Vector
$\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \theta_{ij2}, \theta_{ij3})^T$
$i = 1, \cdots, 51;\ j = 1, \cdots, 10(1989)$

Goal : Find estimates of $\theta_{i\overline{101}}$ $(i = 1, \cdots, 51)$ based on $\boldsymbol{Y}_{ij}$ $(i = 1, \cdots, 51;\ j = 1, \cdots, 10)$ and the associated standard errors.

Auxiliary Information

x_{ij11} (x_{ij12}) = adjusted (base year) census median income of four-person families for state i in year j;
x_{ij21} (x_{ij22}) = adjusted (base year) census median income of three-person families for state i in year j;
x_{ij31} (x_{ij32}) = adjusted (base year) census median income of five-person families for state i in year j.

Write

$\boldsymbol{x}_{ij1} = (1, x_{ij11}, x_{ij12})^T$;
$\boldsymbol{x}_{ij2} = (1, x_{ij21}, x_{ij22})^T$;
$\boldsymbol{x}_{ij3} = (1, x_{ij31}, x_{ij32})^T$.

Let

$$\boldsymbol{X}_{ij} = \begin{pmatrix} \boldsymbol{x}_{ij1}^T & \boldsymbol{0}^T & \boldsymbol{0}^T \\ \boldsymbol{0}^T & \boldsymbol{x}_{ij2}^T & \boldsymbol{0}^T \\ \boldsymbol{0}^T & \boldsymbol{0}^T & \boldsymbol{x}_{ij3}^T \end{pmatrix}$$

$i = 1, \cdots, 51;\ j = 1, \cdots, 10.$

Consider the following HB modeling.
I. $\boldsymbol{Y}_{ij}|\boldsymbol{\theta}_{ij}, \boldsymbol{\alpha}, \boldsymbol{b}_j, \boldsymbol{\Psi}_j, \boldsymbol{W} \sim N(\boldsymbol{\theta}_{ij},\ \boldsymbol{V}_{ij})$;
II. $\boldsymbol{\theta}_{ij}|\boldsymbol{\alpha}, \boldsymbol{b}_j, \boldsymbol{\Psi}_j, \boldsymbol{W} \sim N(\boldsymbol{X}_{ij}\boldsymbol{\alpha} + \boldsymbol{b}_j, \boldsymbol{\Psi}_j)$;
III. $\boldsymbol{b}_j|\boldsymbol{b}_{j-1}, \boldsymbol{W} \sim N(\boldsymbol{b}_{j-1}, \boldsymbol{W})$;
IV. $\boldsymbol{\alpha} \sim$ uniform(R^p);
$\boldsymbol{\Psi}_j^{-1} \overset{ind}{\sim}$ Wishart$(\boldsymbol{S}_j, k_j)$;
$\boldsymbol{W}^{-1} \sim$ Wishart$(\boldsymbol{S}_0, k_0)$
where $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}_j^{-1}$ and $\boldsymbol{W}^{-1}$ are mutually independent.

After attempting several models involving the normal and lognormal distributions, Ghosh, Nangia and Kim (1996) found that there was no need to use both adjusted and base year census median income. In fact, keeping both was making things worse. So the revised design matrix is

$$\boldsymbol{X}_{ij} = \begin{bmatrix} 1 & x_{ij11} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & x_{ij21} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{ij31} \end{bmatrix}$$

$i = 1, \cdots, 51;\ j = 1, \cdots, 10(1989)$

We took $\boldsymbol{S}_0 = \boldsymbol{S}_1 = \cdots = \boldsymbol{S}_{10} = 0.00005\boldsymbol{I}$ and $k_j = 7 \forall j = 0, 1, \cdots, 10$. The results are given the next few tables.

Conclusions :

(1) Time series modeling always results in significant improvement over its non-time series counterpart.

(2) The estimates obtained by the Bureau of the Census have improved significantly now due to Fay's empirical Bayes methodology.

(3) The best results for the HB model seem to come when one considers median income of 4 and 5 person families. The trivariate case gives the next best results. Third in the line is Fay's recommendation of considering $\frac{3}{4}$(median income of 3-person families)+$\frac{1}{4}$(median income of 5-person families).

TABLE 1. COMPARISION OF ESTIMATES

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
HB	0.02074	0.00069	458.73	346085.10
EB	0.02042	0.00068	450.63	334231.14
Bureau	0.03246	0.00165	722.84	835709.94
CPS	0.04984	0.00340	1090.41	1631203.47

TABLE 2. PERCENTAGE IMPROVEMENT OF THE OPTIMAL ESTIMATES OVER OTHERS

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
HB	1.54%	1.45%	1.77%	3.43%
EB	0.0%	0.0%	0.0%	0.0%
Bureau	37.09%	58.79%	37.66%	60.01%
CPS	59.03%	80.00%	58.67%	79.51%

TABLE 3. ESTIMATED STANDARD ERRORS OF ESTIMATES

State	SD.CPS	SD.EB	SD.HB
1	1533	382	601
2	1647	188	515
3	1642	253	554
4	1414	280	566
5	2409	287	591
6	1431	385	606
7	766	440	557
8	1264	358	578
9	878	208	496
10	843	233	465
11	1041	219	482
12	1025	286	507
13	1006	293	509
14	1202	282	523
15	1571	316	566
16	1592	226	531
17	1255	194	498
18	1535	342	591
19	1588	335	575
20	1802	229	544
21	1586	293	553
22	1900	424	664
23	1722	299	575
24	3801	192	545
25	1418	242	532

TABLE 3. (continued)

State	SD.CPS	SD.EB	SD.HB
26	1380	344	575
27	1012	305	517
28	1795	297	563
29	1196	241	548
30	1042	197	518
31	1285	322	548
32	1274	321	546
33	1282	305	566
34	1762	435	647
35	1507	418	620
36	1444	315	567
37	1675	293	569
38	873	274	478
39	1625	262	544
40	1543	253	538
41	1559	671	837
42	1677	364	621
43	1597	297	555
44	1828	194	521
45	1440	201	506
46	2051	254	558
47	1415	254	536
48	1500	246	530
49	767	304	476
50	1891	982	1093
51	2263	355	625

TABLE 4. COMPARISION OF ESTIMATES

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
Bureau	0.03246	0.00165	722.84	835709.94
CPS	0.04984	0.00340	1090.41	1631203.47
EB^1	0.02042	0.00068	450.63	334231.14
HB^1	0.02074	0.00069	458.73	346085.10
HB^{2a}	0.02044	0.00068	452.47	341070.67
HB^{2b}	0.02057	0.00068	454.47	339415.22
HB^{2c}	0.02066	0.00065	455.72	322042.62
HB^3	0.02022	0.00067	447.35	336966.41

TABLE 5. PERCENTAGE IMPROVEMENT OF THE OPTIMAL ESTIMATES OVER OTHERS

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
Bureau	37.71%	60.61%	38.11%	61.46%
CPS	59.43%	80.88%	58.97%	80.26%
EB^1	0.98%	4.41%	0.73%	3.65%
HB^1	2.51%	5.80%	2.48%	6.95%
HB^{2a}	1.08%	4.41%	1.13%	5.58%
HB^{2b}	1.70%	4.41%	1.57%	5.12%
HB^{2c}	2.13%	0.0%	1.84%	0.0%
HB^3	0.0%	2.99%	0.0%	4.63%

HB^1 = univariate with 4-person
HB^{2a} = bivariate with 4-person and 3-person
HB^{2b} = bivariate with 4-person and 5-person
HB^{2c} = bivariate with 4-person and .75*3-person+.25*5-person
HB^3 = trivariate with all 3-person, 4-person and 5-person

NOTATIONS FOR TABLES 6-8:

HB^1 = time uni with 4-person
HB^2 = nontime uni with 4-person
HB^3 = time bi with 4-person and 3-person
HB^4 = nontime bi with 4-person and 3-person
HB^5 = time bi with 4-person and 5-person
HB^6 = nontime bi with 4-person and 5-person
HB^7 = time bi with 4-person and .75*3-person+.25*5-person
HB^8 = nontime bi with 4-person and .75*3-person+.25*5-person
HB^9 = time tri with all 3-person, 4-person and 5-person
HB^{10} = nontime tri with all 3-person, 4-person and 5-person

TABLE 6. COMPARISION OF ESTIMATES

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
Sample	0.0735	0.0084	2928.82	13811122.39
HB^1	0.0338	0.0018	1351.67	3095736.14
	(54.00 %)	(78.18 %)	(53.85 %)	(77.59 %)
HB^2	0.0363	0.0021	1457.47	3468496.61
	(50.61 %)	(75.32 %)	(50.24 %)	(74.89 %)
HB^3	0.0295	0.0013	1173.76	2197974.86
	(59.85 %)	(84.08 %)	(59.92 %)	(84.09 %)
HB^4	0.0322	0.0016	1286.41	2604939.39
	(56.14 %)	(81.02 %)	(56.08 %)	(81.14 %)
HB^5	0.0233	0.0009	943.57	1641855.92
	(68.31 %)	(89.07 %)	(67.78 %)	(88.11 %)
HB^6	0.0295	0.0013	1179.06	2212953.57
	(59.87 %)	(84.46 %)	(59.74 %)	(83.98 %)
HB^7	0.0286	0.0012	1145.37	2095357.61
	(61.06 %)	(85.17 %)	(60.89 %)	(84.83 %)
HB^8	0.0324	0.0015	1295.78	2525149.75
	(55.99 %)	(81.84 %)	(55.76 %)	(81.72 %)
HB^9	0.0274	0.0012	1099.47	1965572.14
	(62.76 %)	(86.14 %)	(62.46 %)	(85.77 %)
HB^{10}	0.0308	0.0014	1235.69	2322768.16
	(58.06 %)	(83.29 %)	(57.81 %)	(83.18 %)

TABLE 7. COMPARISION OF ESTIMATES

Estimate	Avg Rel Bias	Avg Sq Rel Bias	Avg Abs Bias	Avg Sq Dev
Bureau	0.0296	0.0013	1183.90	2151350.18
HB^1	0.0338	0.0018	1351.67	3095736.14
	(-14.19 %)	(-40.48 %)	(-14.17 %)	(-43.90 %)
HB^2	0.0363	0.0021	1457.47	3468496.61
	(-22.60 %)	(-58.94 %)	(-23.11 %)	(-61.22 %)
HB^3	0.0295	0.0013	1173.76	2197974.86
	(0.33 %)	(-2.52 %)	(0.86 %)	(-2.17 %)
HB^4	0.0322	0.0016	1286.41	2604939.39
	(-8.89 %)	(-22.24 %)	(-8.66 %)	(-21.08 %)
HB^5	0.0233	0.0009	943.57	1641855.92
	(21.34 %)	(29.60 %)	(20.30 %)	(23.68 %)
HB^6	0.0295	0.0013	1179.06	2212953.57
	(0.37 %)	(-0.03 %)	(0.41 %)	(-2.86 %)
HB^7	0.0286	0.0012	1145.37	2095357.61
	(3.34 %)	(4.49 %)	(3.25 %)	(2.60 %)
HB^8	0.0324	0.0015	1295.78	2525149.75
	(-9.26 %)	(-16.96 %)	(-9.45 %)	(-17.38 %)
HB^9	0.0274	0.0012	1099.47	1965572.14
	(7.54 %)	(10.76 %)	(7.13 %)	(8.64 %)
HB^{10}	0.0308	0.0014	1235.69	2322768.16
	(-4.11 %)	(-7.62 %)	(-4.37 %)	(-7.97 %)

TABLE 8: BEST ESTIMATES, SIMULATED STANDARD ERRORS AND POSTERIOR VARIANCES

State	HB[5]	SSD	V	V_1	V_2
1	38415	16	2325625	123369	2202256
2	48196	24	2556801	282737	2274064
3	39993	23	2295225	273141	2022084
4	52915	33	2205225	554000	1651225
5	44267	34	3013696	583215	2430481
6	54589	26	2788900	352179	2436721
7	44838	19	1123600	182700	940900
8	54432	25	1677025	319800	1357225
9	40179	11	1081600	65536	1016064
10	41372	37	1909924	680043	1229881
11	38612	16	2070721	121905	1948816
12	42850	13	1290496	89280	1201216
13	43087	17	1214404	147315	1067089
14	40437	24	2064969	298728	1766241
15	42842	19	2292196	186795	2105401
16	36939	24	1838736	286220	1552516
17	38458	15	2298256	107856	2190400
18	33578	39	2295225	757625	1537600
19	32340	25	1646089	307440	1338649
20	36942	36	2427364	661123	1766241
21	37108	32	2656900	522379	2134521
22	42350	19	2496400	182959	2313441
23	50974	29	2989441	429441	2560000
24	41246	21	2825761	223992	2601769
25	45147	21	2576025	219800	2356225

TABLE 8: (continued)

State	HB[5]	SSD	V	V_1	V_2
26	31863	13	1745041	80941	1664100
27	37560	9	1087849	35173	1052676
28	35948	20	2013561	191061	1822500
29	40856	24	2595321	297065	2298256
30	37673	30	1371241	461125	910116
31	33488	19	2114116	179235	1934881
32	34735	35	2442969	606944	1836025
33	34704	20	2105401	201001	1904400
34	31496	13	1874161	83917	1790244
35	31101	19	1836025	177081	1658944
36	33465	17	2190400	142639	2047761
37	32927	23	2241009	266984	1974025
38	35087	33	1633284	543348	1089936
39	33223	31	2229049	478720	1750329
40	32883	14	1590121	89496	1500625
41	35048	46	3143529	1069929	2073600
42	40106	21	2374681	210840	2163841
43	31327	16	2007889	122760	1790244
44	38081	18	2421136	162127	2259009
45	36370	20	2102500	192576	1909924
46	39498	47	2893401	1095120	1798281
47	41216	18	2067844	166203	1901641
48	38277	26	2319529	328608	1990921
49	42845	19	1203409	185328	1018081
50	50191	32	2762244	524228	2238016
51	44265	19	2676496	189567	2486929

REFERENCES

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-compotents model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

Cressie, N. (1992). Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis*, 24.

Datta, G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, 19, 1748-1770.

Datta, G. S., Ghosh, M., Nangia, N. and Natarajan, K. (1995). Estimation of median income of four-person families: A Bayesian approach. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Eds. D.A. Berry et. al. Wiley, New York, pp 129-140.

Dempster, A. P. and Tomberlin, T. J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.

Fay, R. E. (1987). Application of multivariate regression to small domain estimation. *Small Area Statistics*. Eds. R. Platek, N. J. K. Rao, C. E. Särndal, and M. P. Singh. Wiley, New York, 91-102.

Fay, R. E. and Herriot, R. A. (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.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations. *Communications in Statistics-Theory and Methods*, 20(5 & 6),

1747-1766.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, 533-540.

Ghosh, M., Nangia, N. and Kim, D. H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Ghosh, M., Natarajan, K., Stroud, T. W. K., and Carlin, B. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., Natarajan, K., Waller, L. A., and Kim, D. H. (1999). Hierarchical Bayes GLM's for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

Holt, D., Smith, T. M. F. and Tomberlin, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

Kim, D. H. (1998). Bayesian robustness in small area estimation. *Statistics & Decision*, 16, 89-103.

Kim, D. H. (2002). Bayesian and empirical Bayesian analysis under informative sampling. To appear in *Sankhya*.

Lahiri, P. and Rao, N. J. K. (1994). Robust estimation of mean square error of small area estimators. To appear in *Journal of the American Statistical Association*.

MacGibbon, B. and Tomberlin, T. J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.

Malec, D., Sedransk, J. and Tompkins, L. (1991). Bayesian predictive inference for small areas for binary variables in the national health interview survey.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association*, 78, 47-65.

Tsutakawa, R. K. (1985). Estimation of cancer mortality rates: A Bayesian analysis of small frequencies. *Biometrics*, 41, 69-79.