INDIRECT ESTIMATION BASED ON ADMINISTRATIVE RECORDS AND THE AMERICAN COMMUNITY SURVEY

Nanak Chand , Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, D.C. 20233

1. Introduction

The American Community Survey (ACS) consists of monthly rolling samples designed to update annually the social and economic profile that the U.S. census traditionally provided once a decade. While the ultimate ACS sampling rate will be about three percent of the population in most areas, the corresponding rate for the 1996 demonstration sites is fifteen percent for majority of the larger units and thirty percent for the smaller units. The demonstration sites are Brevard County Florida, Fulton County Pennsylvania, Multnomah County/Portland Oregon, and Rockland County New York. Though based on different definitions suitable for their initial objectives, the household surveys, the census, and the administrative records are the main sources of data for policy and planning.

The ACS sample size is designed to result in reliable direct estimates for substate areas. For small areas, such as census tracts, it is desirable to improve the ACS estimates by borrowing strength from neighboring areas and other sources of data. This paper develops indirect estimates o f characteristics of interest by integrating 1996 ACS data with the Internal Revenue Service records.

The resulting estimates are composite of the direct and synthetic regression estimators based on random area effect models (Chand and Alexander (1995), Cressie (1989, 1990, 1992), Datta et al (1992), Ericksen and Kadane (1985, 1987, 1992), Fay (1987), Fay and Herriot (1979), Ghosh and Rao (1994), Prasad and Rao (1990), Singh, Gambino and Mantel (1994), and Spjotvoll and Thomsen (1987).)

Subsequent sections describe the model and underlying assumptions, depict different methods of estimating the variance components, derive empirical Bayes estimators along with appropriate measures of precision, and define a class of modified estimators.

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

The paper also illustrates the methods by developing estimators of tract level poverty rates for three of the 1996 ACS sites, provides measures of reduction in variance achieved by the procedure, and gives additional test statistics as well as comparisons of estimators produced by the different methods. Fulton county has been excluded from analysis due to small number of tracts in the sample.

2. Assumed Model and the Estimation of Variance Components

A large area A is composed of m small areas A_i , i = 1, ..., m. The parameter of interest for A_i is the true population proportion P_i . A direct estimator p_i of P_i is available from the ACS. The auxiliary data $\underline{x}_i = (x_{i1}, \ldots, x_{is})^T$ are available from the administrative records for each A_i .

The transformation g is a function of a single variable and has a nonzero continuous first derivative. Let

$$g_i = g(p_i), \; i = 1, \ldots, m. \; .$$

We consider the small area model,

$$\underline{g} = X\underline{\beta} + \underline{t} + \underline{e} \, ,$$

where $\underline{g}$, $\underline{t}$, and $\underline{e}$ are m x 1 vectors, $\underline{e}$ represents random sampling errors, $\underline{t}$ represents random area effects, and $\underline{g}$ has a multivariate normal distribution. X is a mxs design matrix, $\underline{\beta}$ is a sx1 vector of unknown parameters, and $\underline{t}$ and $\underline{e}$ are statistically independent. Let $\sum$ and ∇ be mxm diagonal matrices with the (i,i)th elements respectively equal to τ^2 and δ_i^2.

We also assume that

$$E(\underline{e} \mid \underline{g}) = \underline{0}, \; Var(\underline{e} \mid \underline{g}) = \nabla,$$

and $\underline{t} \sim N(\underline{0}, \sum)$.

In this paper, we study the variance stabilization transformation given by

$$g_i = 2\sin^{-1}(\sqrt{p_i}) \;, \quad i = 1,\ldots, m.$$

(Cox and Snell (1989)). The suitability of the above assumptions under this transformat i o n is tested in a later section.

We consider four estimators of the variance component τ^2 under the above model.. These are the maximum likelihood (ML) estimator, the restricted maximum likelihood (RML) estimator (Cressie (1989, 1992)), the Fay Herriot (FH) estimator (Fay and Herriot (1979)), and a quadratic moment (QM) estimator (Prasad and Rao (1990) and Ghosh and Rao (1994).)

The ML estimators of $\underline{\beta}$ and τ^2 minimize the expression

$$\ln(|V|) + (\underline{g} - X\underline{\beta})^T V^{-1}(\underline{g} - X\underline{\beta})$$

where V is a mxm diagonal matrix with the (i, i)th element equal to $\tau^2 + \delta_i^2$.

The asymptotic variance of $\hat{\tau}^2$ (ML) is given by

$$V(ML) = [\frac{1}{2}\sum_{i=1}^{m} (\delta_i^2 + \tau^2)^{-2}]^{-1}.$$

The RML estimators of $\underline{\beta}$ and τ^2 minimize

$$\ln(|V|) + \ln(|X^T V^{-1} X|)$$

$$+ (\underline{g} - X\underline{\beta})^T V^{-1} (\underline{g} - X\underline{\beta}).$$

The asymptotic variance of RML estimator of τ^2 is given by

$$V(RML) = [\frac{1}{2} trace(\pi(\tau^2)\pi(\tau^2))]^{-1}, \text{ with}$$

$$\pi(\tau^2) = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}.$$

The FH estimator of τ^2 is obtained by simultaneously solving

$$(\underline{g} - X\underline{\beta}))^T V^{-1}(\underline{g} - X\underline{\beta}) = m-s,$$

and

$$\underline{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \underline{g}$$

The QM estimator of τ^2 is given by

$$(m-s)^{-1}[(\underline{g} - X\underline{\hat{b}})^T(\underline{g} - X\underline{\hat{b}}) - \sum_{i=1}^{m} \delta_i^2 + \sum_{i=1}^{m} \delta_i^2 \underline{x}_i^T (X^T X)^{-1} \underline{x}_i]$$

where $\hat{b}$ is the ordinary least square estimator of $\underline{\beta}$ given by

$$\underline{\hat{b}} = (X^T X)^{-1} X^T \underline{g},$$

and $\underline{x}_i^T$ is the ith row of the design matrix X. Under normality, the variances of FH and QM estimators of τ^2 are

$$V(FH) = V(QM) = 2m^{-2}\sum_{i=1}^{m}(\delta_i^2 + \tau^2)^2$$

3. Empirical Bayes (EB) Estimators, their Precision, and the Modified Small Area Estimators

With τ^2 estimated by one of the four methods in section 2, let $\hat{\underline{\beta}}$ be the best linear unbiased estimator of $\underline{\beta}$ given by

$$\hat{\underline{\beta}} = (X^T U^{-1} X)^{-1} X^T U^{-1} \underline{g} ,$$

where U is the mxm matrix obtained from V by replacing τ^2 by its estimator $\hat{\tau}^2$. Let

$$\gamma_i = \tau^2 / (\tau^2 + \delta_i^2) ,$$

be the measure of uncertainty in the model relative to the total variance. Then the regression synthetic estimator of $g(P_i)$ is $X^T\hat{\underline{\beta}}$ and the EB of $g(P_i)$ is given by

$$\hat{g}_i = \hat{\gamma}_i g_i + (1-\hat{\gamma}_i)\underline{x}_i^T\hat{\underline{\beta}}$$

where $\hat{\gamma}_i$ is the value of γ_i when τ^2 is replaced by its estimator $\hat{\tau}^2$. The corresponding estimator of P_i is obtained by inverting g.

The MSE of $\hat{g}_i$ (Cressie (1992), Kacker and Harville (1984), and Ghosh and Rao (1994)) is given by

$$M_i^g = M_{0i}(\tau^2) + \delta_i^4(\tau^2 + \delta_i^2)^{-3} v^a(\hat{\tau}^2) ,$$

where $v^a(\hat{\tau}^2)$ is the asymptotic variance of $\hat{\tau}^2$ and

$$M_{0i}(\tau^2) = \gamma_i \delta_i^2 + (1-\gamma_i)^2 \underline{x}_i^T (X^T V^{-1} X)^{-1} \underline{x}_i .$$

Since ACS is designed to provide unbiased estimates for large areas, we make an adjustment to the above estimator,by taking the modified estimator as the sum of the EB estimator for a specific area and a predetermined weight times the the difference between the direct survey estimate and the weighted average of the EB estimators for each of the small areas.

4. Estimation of Proportion of Persons Below Poverty Level

We illustrate the above estimation procedures by taking $\{ A_i, i = 1, \ldots, m \}$ as the census tracts respectively in Brevard County Florida, Multnomah County/Portland Oregon, and Rockland County New York.

The direct estimate p_i of the proportion below poverty level in A_i is calculated as the ratio of weighted number of persons below poverty level to the total weighted ACS population in the respective tract. The function g is taken as described in section 2.

The design matrix X is defined with s = 6 based on the Internal Revenue Service variables as

$$X_{i1} = 1,\ X_{i2} = \ln\,[\mathit{Median\ Income}]$$

$$X_{i3} = \ln\,[\ \mathit{Per\ Capita\ Income}\ ] ,$$

$$X_{i4} = \ln\,[\ Q_L\ ] ,$$

$$X_{i5} = \ln\,[\ Q_U\ ] ,$$

and

$$X_{i6} = 2\sin^{-1}\sqrt{P_V} \ ,$$

where Q_L, Q_U, and P_V are respectively, the lower quartile income, upper quartile income, and proportion of persons below poverty level in the tract.

We tested the appropriateness of the models by veryfing that the standardized residuals given by

$$r_i = (g_i - \underline{x}_i^T\hat{\beta})/\sqrt{(\hat{\tau}^2 + \delta_i^2)},$$

i = 1, ..., m, are approximately normally distributed with mean zero and variance one.

5. A Comparison of the Variance Component Methods

Table A shows the four sets of EB estimators of proportions below poverty level for randomly selected tracts for one of the three sites. There are small differences among the four sets of estimated values.

Table B shows the modified EB estimators of proportions below poverty level. An appropriately weighted sum of these estimators equals the ACS estimate of proportion below poverty level for the whole county. Tables C gives MSE estimates associated with the four EB estimators. The table shows the small levels of MSE of the EB estimators for each of the estimation methods.

6. Analysis Applicable to the Ultimate ACS Size Levels

Since the ultimate ACS sample will be about twenty percent of the 1996 sample, we perform the following analysis appropriate for the ultimate size levels. For area i, let $p_i^{(k)}$ denote the direct estimate of proportion of persons in poverty in the kth systematic sample of one-fifth size taken from the full ACS sample for a specified site, and let $p_i^{(ck)}$ denote the corresponding estimate from the remaining four-fifth sample, i = 1, ..., m; k = 1, ..., 5. Also, let $g_i^{(k)}$ and $g_i^{(ck)}$ be the corresponding transformed values. We repeat the analysis of sections 2 - 4 replacing p_i by $p_i^{(k)}$, i = 1, ..., m ; k= 1, ..., 5.

Let $\hat{g}_i^{(k)}$ and $\hat{p}_i^{(k)}$ be the kth sample estimators derived similar to the full sample case, and let $\hat{M}_i^{g(k)}$ and $\hat{M}_i^{(k)}$, be the corresponding estimates of the their mean squared errors. Also let

$\hat{V}_i^{g(ck)}$ and $\hat{V}_i^{(ck)}$ be the variance estimates of $g_i^{(ck)}$ and $p_i^{(ck)}$ respectively. Then we study the following 2m test statistics:

$$s_i^g = \frac{\hat{g}_i^{(k)} - g_i^{(ck)}}{\sqrt{\hat{M}_i^{g(k)} + \hat{V}_i^{g(ck)}}} \ , i = 1, ..., m, \text{and}$$

$$s_i = \frac{\hat{p}_i^{(k)} - p_i^{(ck)}}{\sqrt{\hat{M}_i^{(k)} + \hat{V}_i^{(ck)}}} \ , i = 1, ..., m.$$

These statistics provide a measure to test the difference between the model estimators given by the one-fifth sample as compared with the larger complementary four-fifth sample estimates, for each of the m areas. Table D gives values of s_i^g and s_i for five of the randomly selected areas for Brevard county. The overall reduction in variance produced by the procedure is given by the following table:

Site	m	ACS Variance x1000	MSE x1000	Percent Reduction
Brevard	86	0.4727	0.3065	35.16%
Multnomah	164	1.0728	0.3775	64.81%
Rockland	39	0.1401	0.1129	19.41%
Composite		0.5619	0.2656	52.73%

Tables A -D for Brevard county follow. The reference list is available from the authors.

Table A
ESTIMATES (EB) OF 1996 POVERTY RATES
Brevard County, Florida

Tract	RML	ML	FH	QM
60100	0.18646	0.18591	0.18630	0.18605
60900	0.24629	0.22237	0.22340	0.22273
64500	0.14801	0.14805	0.14802	0.14804
65232	0.09276	0.09340	0.09293	0.09324
66600	0.05002	0.05014	0.05005	0.05011

Table B

MODIFIED ESTIMATES OF 1996 POVERTY RATE
Brevard County, Florida

Tract	RML	ML	FH	QM
60100	0.18882	0.18844	0.18872	0.18854
60900	0.22664	0.22529	0.22627	0.22562
64500	0.14976	0.14993	0.14980	0.14989
65232	0.09386	0.09459	0.09406	0.09441
66600	0.05053	0.05069	0.05057	0.05065

Table C

MEAN SQUARE ERRORS OF ESTIMATES OF
1996 POVERTY RATES
Brevard County, Florida

Tract	RML	ML	FH	QM
60100	.00028239	.0002774	.00028129	.00027904
60900	.00060523	.00058524	.00060050	.00059124
64500	.00027347	.00026871	.00027244	.00027028
65232	.00008744	.00008724	.00008743	.00008736
66600	.00007784	.00007697	.00007766	.00007728

Table D

TEST STATISTICS FOR SAMPLE k FOR THE 1996
POVERTY RATES
Brevard County, Florida
k= 1

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
60100	-1.74770	-1.83082	-1.76403	-1.84626
60900	-0.85444	-0.87546	-0.86870	-0.88910
64500	-0.63460	-0.64087	-0.64473	-0.65096
65232	-1.18542	-1.22540	-1.11329	-1.14782
66600	1.58199	1.45221	1.55936	1.43529

Table D (Continued)

TEST STATISTICS FOR SAMPLE k FOR THE 1996
POVERTY RATES
Brevard County, Florida
k=1

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
60100	-1.74219	-1.82561	-1.73817	-1.82182
60900	-0.84986	-0.87109	-0.84638	-0.86778
64500	-0.63144	-0.63773	-0.62881	-0.63510
65232	-1.20663	-1.24835	-1.22657	-1.26990
66600	1.58811	1.45660	1.59463	1.46152

Table D

TEST STATISTICS FOR SAMPLE k FOR THE 1996
POVERTY RATES
Brevard County, Florida
k= 2

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
60100	0.54730	0.54020	0.47949	0.47423
60900	-0.33051	-0.33267	-0.42354	-0.42680
64500	-0.32995	-0.33426	-0.33095	-0.33502
65022	-0.03524	-0.03527	0.04106	0.04101
66600	-1.81682	-1.92447	-1.79178	-1.89121

Table D (Continued)

TEST STATISTICS FOR SAMPLE k FOR THE 1996
POVERTY RATES
Brevard County, Florida
k=2

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
60100	0.57838	0.57028	0.60374	0.59483
60900	-0.28481	-0.28649	-0.24998	-0.25131
64500	-0.32892	-0.33335	-0.32915	-0.33368
65022	-0.07199	-0.07215	-0.10067	-0.10100
66600	-1.82613	-1.93841	-1.83607	-1.95133

Table D

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k= 3

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
60100	0.64074	0.63299	0.60457	0.59788
60900	-1.52629	-1.57597	-1.58337	-1.63247
64500	-1.61923	-1.71892	-1.60776	-1.70041
65232	1.70914	1.62623	1.73548	1.65170
66600	1.77459	1.65148	1.76978	1.65067

Table D (Continued)

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k=3

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
60100	0.65092	0.64281	0.66025	0.65184
60900	-1.50625	-1.55619	-1.49200	-1.54196
64500	-1.62035	-1.72245	-1.62416	-1.72816
65232	1.69952	1.61685	1.69322	1.61079
66600	1.77411	1.64963	1.77572	1.65027

Table D

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k= 4

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
60100	-2.06925	-2.13982	-2.09032	-2.15872
60900	-1.15603	-1.16437	-1.23296	-1.24080
64500	0.43759	0.43540	0.42207	0.42017
65232	0.56598	0.55957	0.60512	0.59799
66600	-1.53890	-1.62913	-1.51955	-1.60319

Table D (Continued)

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k=4

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
60100	-2.06707	-2.13771	-2.08528	-2.15400
60900	-1.17809	-1.18674	-1.25776	-1.26603
64500	0.43043	0.42829	0.41278	0.41093
65232	0.57801	0.57131	0.61910	0.61159
66600	-1.52529	-1.61428	-1.50145	-1.58402

Table D

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k= 5

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
60100	0.28949	0.28834	0.27141	0.27041
60900	0.31931	0.31719	0.27346	0.27196
64500	1.94839	1.83509	1.93758	1.82760
65232	-2.45572	-2.58874	-2.41874	-2.54602
66600	-0.53039	-0.55661	-0.52151	-0.54622

Table D (Continued)

TEST STATISTICS FOR SAMPLE k FOR THE 1996 POVERTY RATES
Brevard County, Florida
k=5

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
60100	-0.03016	-0.03017	0.28393	0.28282
60900	-0.38549	-0.38696	0.30519	0.30327
64500	1.72982	1.66455	1.94326	1.83093
65232	-1.81925	-1.87870	-2.44383	-2.57522
66600	-0.41799	-0.42907	-0.52713	-0.55289