Applying Bivariate Binomial/Logit Normal Models to Small Area Estimation

Carolina Franco[1] and William R. Bell[2]

[1]Center for Statistical Research and Methodology, [2]Research and Methodology Directorate
U.S. Census Bureau
Washington, DC 20233

Abstract

The U.S. Census Bureau's SAIPE (Small Area Income and Poverty Estimates) program estimates poverty for various age groups for states, counties, and school districts of the U.S. We focus here on poverty estimates of school-aged (5-17) children for counties. The corresponding SAIPE production model applies to log transformed direct survey estimates for each county of the number of 5-17 year-olds in poverty, with logged covariates obtained from tabulations of administrative record sources (e.g., tax return data) and a previous census (2000) long form estimate. We explore an alternative model assuming a binomial distribution for rescaled survey estimates of the number of school-aged children in poverty, with an effective sample size defined so the variances of the binomial proportions equal the corresponding sampling variance estimates. The model assumes a normal distribution for logits of the underlying county poverty rates, with a mean function using logit covariates derived from the covariates of the production SAIPE model, and with additive random effects. We apply a bivariate version of this model to direct estimates from the American Community Survey (ACS) for 2011 and the previous 5-year (2006-2010) ACS estimates.

Keywords: Small Area Estimation; Complex Surveys; American Community Survey; Bivariate Model; SAIPE.

1. Introduction

The U.S. Census Bureau's SAIPE (Small Area Income and Poverty Estimates) program provides updated poverty estimates for various age groups for states, counties, and school districts of the U.S. These estimates are provided to facilitate the administration of federal programs and the allocation of federal funds to local jurisdictions. This includes the use of SAIPE's school district age 5-17 poverty estimates by the U.S. Department of Education in allocating funds (over $14 billion in 2011) under Title I of the No Child Left Behind Act of 2001.

SAIPE produces state and county poverty estimates using linear Fay-Herriot models (Fay and Herriot 1979, Rao 2003) applied to direct poverty estimates from Census Bureau surveys. Prior to 2005, SAIPE used estimates from the Current Population Survey (CPS) for this purpose; since 2005, SAIPE has used estimates from the American Community Survey (ACS) due to its larger sample size. Covariates in the SAIPE Fay-Herriot models come from tabulations of administrative records data related to income and poverty (tax data and Supplemental Nutrition Assistance Program (SNAP) data), plus poverty estimates from the previous Decennial Census. Since there was no long form sample in the 2010 Census, and hence no 2010 Census poverty estimates, the "previous Census estimates" since 2000 have been those from Census 2000. Due to lack of reliable updated covariate data at the school district level, SAIPE has not yet used formal statistical models in producing the school district poverty estimates. SAIPE has instead relied on an allocation scheme applied to the model-based county poverty estimates that makes use of both previous census poverty estimates and current year tabulations of tax data for school districts (Maples and Bell 2007). Further information on SAIPE models, estimation procedures, and data sources

is given on the SAIPE web site at http://www.census.gov/did/www/saipe/index.html. Further information on the ACS is available at http://www.census.gov/acs/www/.

In this paper, we focus on modeling county poverty rates of school-aged (5-17) children. In particular, we examine a model that assumes a binomial distribution for "observed" sample numbers of school-aged children in poverty by counties. The observed values are obtained by multiplying the direct ACS estimate of the county poverty rate for school-age children by an "effective sample size" determined so the resulting binomial variance reproduces an estimate of the variance of the direct ACS estimate of the number of 5-17 year-old children in poverty. The true binomial proportion p_i (true 5-17 poverty rate) is assumed to follow a linear model for $\log(p_i/(1-p_i))$.

Our model features another change to the SAIPE production county model, this in regard to the use of previous Census estimates. Though the estimates from Census 2000 still provide useful covariates (having statistically significant coefficients) in the state and county models, there are natural concerns about possible declines in their relevance as they become further out-of-date. Since the ACS has now supplanted the census long form, it is also natural to consider ways of using recent past ACS estimates to replace the previous census estimates in the SAIPE models. Huang and Bell (2012) studied this for linear bivariate Fay-Herriot models of state and county level poverty and found no indication that replacing Census 2000 estimates by prior ACS 5-year estimates would lead to a loss of accuracy, and found some suggestion that this could produce a slight improvement. Further improvements could be realized as time progresses and the Census 2000 estimates become more distant from the current production year.

We bring ACS 5-year estimates into our model by using a bivariate binomial/logit normal model analogous to a linear bivariate Fay-Herriot model. (For discussion of the latter, see Bell (2000) or Huang and Bell (2012).) The important aspect of the bivariate model is that it appropriately recognizes the varying levels of sampling error across counties in the ACS 5-year estimates. This aspect is ignored if prior ACS estimates are simply included in the model as an additional regressor.[1] Huang and Bell (2012, pp. 18-22) illustrate this point.

The rest of the paper proceeds as follows. Section 2 reviews the production SAIPE county 5-17 poverty model. Section 3 presents the bivariate binomial / logit normal model, while Section 4 presents results from application of this model to ACS data for 2011, and comparisons of these results to results from the production county model. Section 5 considers future research. Some additional alternative bivariate county poverty models are discussed there, some of whose univariate versions have been previously investigated by some of our Census Bureau colleagues. These additional alternatives include some nonlinear Gaussian models and some models with alternative (to the logit) link functions. We also discuss a model that, rather than use past ACS 5-year estimates, incorporates 5 years of past ACS single-year estimates individually via a first-order autoregressive structure.

[1]This issue also exists with using previous census estimates as a regression variable in county models, though it is less important there since the census long form had an even larger sample than that of ACS. The issue does not arise for using previous census estimates in SAIPE state models, since at the state level the long form estimates had negligible sampling error.

2. The SAIPE County 5-17 Poverty Model

The production SAIPE model follows the general model of Fay and Herriot (1979), which can be written

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m \tag{1}$$

$$Y_i = \mathbf{x}_i'\beta + u_i \tag{2}$$

where Y_i is the population characteristic of interest for area i, y_i is the direct survey estimate of Y_i, e_i is the sampling error in y_i, $\mathbf{x}_i$ is a $p \times 1$ vector of values of regression variables for area i, β is the corresponding vector of regression parameters, and m is the number of small areas. The sampling errors e_i are generally assumed to be distributed independently over i as $N(0, v_i)$. The sampling variances v_i are typically treated as known although, in reality, they are estimated from survey microdata. The area random effects u_i (also called "model errors") are usually assumed to be distributed *i.i.d.* $N(0, \sigma_u^2)$ and independently of the e_i.

The unknown parameters of the model given by (1)–(2) are β and σ_u^2. Given a value for σ_u^2, β can be estimated by weighted least squares regression of y_i on $\mathbf{x}_i$ for $i = 1, \ldots, m$ using weights $(\sigma_u^2 + v_i)^{-1}$. There are various ways of estimating σ_u^2 such as method of moments, maximum likelihood (ML), or restricted maximum likelihood (REML). One can apply these in an iteration that alternates estimation of β given σ_u^2 with estimation of σ_u^2 given β.

If σ_u^2 and the v_i are known, standard best linear unbiased prediction (BLUP) results give the predictors $\hat{Y}_i$ and their error variances:

$$\widehat{Y}_i = h_i y_i + (1 - h_i)\mathbf{x}_i'\widehat{\beta} \tag{3}$$

$$\text{Var}(Y_i - \widehat{Y}_i) = \sigma_u^2(1 - h_i) + (1 - h_i)^2\mathbf{x}_i'\text{Var}\left(\widehat{\beta}\right)\mathbf{x}_i \tag{4}$$

where $h_i = \sigma_u^2/(\sigma_u^2 + v_i)$. In practice, we substitute $\hat{\sigma}_u^2$ and the estimated v_i into the above formulas and into the expression for h_i. Asymptotic corrections to $\text{Var}(Y_i - \widehat{Y}_i)$ are available to account for error in the estimation of σ_u^2. One can also use a Bayesian approach to develop posterior distributions of the model parameters given the data $\mathbf{y} = (y_1, \ldots, y_m)'$, as well as posteriors of the unobserved population characteristics Y_i. The latter provide Bayesian prediction results analogous to (3) and (4) through the posterior means and variances, $E(Y_i|\mathbf{y})$ and $\text{Var}(Y_i|\mathbf{y})$. See Rao (2003) for more details on estimation and prediction for the Fay-Herriot model.

The SAIPE production county 5-17 poverty model is of the form of (1) and (2) with a transformation, namely, y_i equals the logarithm of the ACS estimate of the number of persons age 5-17 in poverty for county i, and Y_i equals the logarithm of the true number of persons age 5-17 in poverty in the county. The sampling variances of the logged survey estimates are estimated directly by a replication method that is discussed in the context of the SAIPE state model by Fay and Train (1995). The unknown model parameters β and σ_u^2 are estimated by ML. Prediction results from (3) and (4) apply on the log scale, and are translated to prediction results for the number of 5-17 year-olds in poverty using properties of the lognormal distribution. Thus, the predictor of the number of 5-17 year-olds in poverty is $\exp\{\widehat{Y}_i + \frac{1}{2}\text{Var}(Y_i - \widehat{Y}_i)\}$, and the estimated prediction MSE is $\exp\{\text{Var}(Y_i - \widehat{Y}_i) - 1\}\exp\{2\widehat{Y}_i + \text{Var}(Y_i - \widehat{Y}_i)\}$. The predictions are then raked (rescaled) to force agreement with corresponding state level predictions of the number of 5-17 year-olds in poverty obtained from the SAIPE state model. Adjustments are also made to the county prediction error variances to approximately account for this raking.

The regression variables in $\mathbf{x}_i$ for the SAIPE 5-17 county poverty model include the constant 1 for an intercept term, and the following:

- log of the number of "poor child exemptions" for the county, i.e., child exemptions claimed on tax returns whose adjusted gross income falls below the official poverty threshold for a family of the size implied by the number of exemptions on the form;
- log of the number of county SNAP benefits recipients in July of the previous year;
- log of the estimated county population age 0-17 as of July 1;
- log of the total number of child exemptions in the county claimed on tax returns; and
- log of the Census 2000 county estimate of the number of related children in poverty ages 5 to 17.

"SNAP" refers to the Supplemental Nutrition Assistance Program (formerly known as the Food Stamp Program) managed by the Food and Nutrition Service of the U.S. Department of Agriculture, which provides the recipients' data.

For some counties with small samples, the direct ACS estimate of the number of 5-17 year-olds in poverty is zero. Since logs cannot be taken of these zero estimates, such counties are dropped from the model fitting. The number of counties dropped varies some year-to-year, but is always small with ACS data. In 2011, for example, 126 counties were dropped out of 3,143 counties in the U.S. Also, since the counties dropped invariably have small samples, the proportion of the sample information dropped is even less than is indicated by the number of counties dropped and so has very little effect on the model fitting and predictions. Prediction results for the counties dropped can still be obtained with the estimated model from (3) and (4), treating such counties as if they supplied no data. One can equivalently look at this as letting $v_i \to \infty$ for any county with zero poor 5-17 in sample, implying $h_i \to 0$ and $\widehat{y}_i \to \mathbf{x}_i'\widehat{\beta}$ in (3).

The preceding discussion referred to the county 5-17 poverty model used for the 2011 estimates. Model changes from year-to-year, when any changes at all are made, have generally been slight, except in 2005 with the switch from using CPS to using ACS data. Further information on the SAIPE models, including documentation of the models used in prior years and discussion of the input data sources, can be found on the SAIPE web site at http://www.census.gov/did/www/saipe/.

3. A Bivariate Binomial/Logit Normal Model for County 5-17 Poverty

The linear Fay-Herriot model may be replaced by a Generalized Linear Mixed Model (GLMM) as discussed by Ghosh, et al. (1998) and Rao (2003, sections 5.6 and 10.11). This may be done when the observed data y_i are inherently discrete, as when they are (unweighted) counts of sampled persons or households with certain characteristics. One such model assumes a Binomial distribution for y_i with success probability p_i, and a logistic regression model for p_i with Gaussian errors in the logit scale. The resulting model is

$$y_i|p_i, n_i \sim \text{Bin}(n_i, p_i) \qquad i = 1, \ldots, m \tag{5}$$

$$\text{logit}(p_i) = \mathbf{x}_i'\beta + u_i \tag{6}$$

where $\text{logit}(p_i) = \log[p_i/(1-p_i)]$, $u_i \sim N(0, \sigma_u^2)$, and n_i is the sample size for area i.

Slud (2000, 2004) did several analyses comparing results from GLMM models to results from models similar to the SAIPE county production model (in the form applied in

earlier years to CPS data). Slud (2000) showed advantages to use of a unit level binomial/logit normal model compared to a linear Fay-Herriot model for logged estimates of number in poverty when the data were simulated from the GLMM model. Simulation settings were chosen to reflect actual SAIPE data. Slud (2004) compared variants of both the GLMM model and the SAIPE production model in fits to CPS data and in predictions to Census 2000 long form estimates, finding the GLMM models fit the CPS data better while the production model better predicted the Census 2000 estimates.

The model given by (5)–(6) can be readily applied to unweighted sample counts y_i, but this ignores any complex aspects of the survey design. In applications to complex survey data where the y_i are weighted estimates, two problems arise. First, the possible values for the y_i will not be the integers $0, 1, \ldots, n_i$ for any direct definition of sample size n_i. Instead, y_i will take a value from a finite set of unequally-spaced numbers determined by the survey weights that apply to the sample cases in area i. Second, the sampling variance of y_i implied by the Binomial distribution in (5), $n_i p_i(1-p_i)$, will be incorrect. We address these two problems by defining an "effective sample size" $\tilde{n}_i$, and an "effective sample number of successes" $\tilde{y}_i$ determined to maintain: (i) the direct survey estimate $\hat{p}_i$, of the poverty proportion, i.e., $\tilde{y}_i/\tilde{n}_i$; and (ii) a corresponding sampling variance estimate, $\widehat{\text{Var}}(\hat{p}_i)$.

We set

$$\tilde{n}_i = \breve{p}_i(1-\breve{p}_i) \Big/ \widehat{\text{Var}}(\hat{p}_i) \tag{7}$$

where $\breve{p}_i$ is a preliminary model-based prediction of the population proportion p_i (on which $\text{Var}(\hat{p}_i)$ truly depends), and $\widehat{\text{Var}}(\hat{p}_i)$ depends on $\breve{p}_i$ through a fitted generalized variance function (GVF). The GVF is discussed in Section 4 and detailed instructions for its implementation are included in the Appendix. We then set $\tilde{y}_i = \tilde{n}_i \times \hat{p}_i$ and, after rounding, substitute $(\tilde{n}_i, \tilde{y}_i)$ for (n_i, y_i) in (5). Note that $\tilde{y}_i = 0$ if $\hat{p}_i = 0$, but this does not cause problems since $\breve{p}_i > 0$ in (7) implies $\tilde{n}_i > 0$. Rounding of $\tilde{n}_i$ and $\tilde{y}_i$ will likely be required by most computer software for the fitting of models such as (5)–(6). Liu, Lahiri, and Kalton (2007) and You (2008) used essentially this sampling variance model, but applied it in models of directly estimated survey proportions assumed to follow either a normal or a Beta distribution.

We shall extend the model given by (5)–(6) to a bivariate version, written as

$$\tilde{y}_{1i}|p_{1i}, \tilde{n}_{1i} \sim \text{Bin}(\tilde{n}_{1i}, p_{1i}) \qquad \tilde{y}_{2i}|p_{2i}, \tilde{n}_{2i} \sim \text{Bin}(\tilde{n}_{2i}, p_{2i}) \tag{8}$$

$$\text{logit}(p_{1i}) = \mathbf{x}'_{1i}\beta_1 + u_{1i} \qquad \text{logit}(p_{2i}) = \mathbf{x}'_{2i}\beta_2 + u_{2i} \tag{9}$$

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim i.i.d.\ N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

for $i = 1, \ldots, m$. In our applications of this model, the $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ are the (effective) sample numbers of 5-17 in poverty from the current ACS 1-year estimates and the prior ACS 5-year estimates, respectively, and, $\tilde{n}_{1i}$ and $\tilde{n}_{2i}$ are the corresponding effective sample sizes, determined as in (7). Note that $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ are assumed conditionally independent (given $p_{1i}, \tilde{n}_{1i}$ and $p_{2i}, \tilde{n}_{2i}$) since the ACS samples are drawn approximately independently each year.

We estimate the model given in (8)–(9) by ML using the NLMIXED procedure of SAS (SAS Institute Inc. 2010). The likelihood function to be maximized is proportional to

$$L(\beta_1, \beta_2, \Sigma) = \prod_{i=1}^{m} \int \int p(\tilde{y}_{1i}|p_{1i}, \tilde{n}_{1i}) p(\tilde{y}_{2i}|p_{2i}, \tilde{n}_{2i}) p(u_{1i}, u_{2i}|\Sigma) du_{1i} du_{2i} \tag{10}$$

where $p(\tilde{y}_{1i}|p_{1i}, \tilde{n}_{1i}) \propto p_{1i}^{\tilde{y}_{1i}}(1-p_{1i})^{\tilde{n}_{1i}-\tilde{y}_{1i}}$ with $p_{1i} = \exp(\mathbf{x}'_{1i}\beta_1 + u_{1i})/[1+\exp(\mathbf{x}'_{1i}\beta_1 + u_{1i})]$, and similarly for $p(\tilde{y}_{2i}|p_{2i}, \tilde{n}_{2i})$, and $p(u_{1i}, u_{2i}|\Sigma)$ is the bivariate normal density with

mean zero and covariance matrix Σ. NLMIXED approximates the integral in (10) by adaptive Gaussian quadrature, and then numerically maximizes the approximated likelihood. Predictions of u_{1i} and u_{2i} in (9) are obtained by setting them to their "conditional modes," $\hat{u}_{1i}$ and $\hat{u}_{2i}$, where these maximize the contribution to the likelihood from area i, this being the integrand in (10). Predictions of the population targets p_{1i}, the county population 5-17 poverty rates for the current year, are then obtained by substituting $\hat{u}_{1i}$ and the MLE of β_1 into $p_{1i} = \exp(\mathbf{x}'_{1i}\beta_1 + u_{1i})/[1 + \exp(\mathbf{x}'_{1i}\beta_1 + u_{1i})]$. The analogous calculation will predict p_{2i}. Prediction variances are obtained by linearization of the predictors. See the NLMIXED documentation for details.

When the sampling design is not complex, generalized linear model theory guarantees the consistency and asymptotic normality of the model estimators (see, for instance, McCulloch and Searle, 2001). However, clearly we have a deviation from standard theory, which we attempt to alleviate through use of the design effect. Especially in small domains, the design effect is a heuristic way to capture the complex aspects of the survey. However, as the area sample size grows, the variance approximation based on the design effect becomes more accurate.

Very similar prediction results were obtained from a Bayesian approach with flat priors on β_1, β_2, σ_{11}, σ_{22}, and $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ using the JAGS software (Plummer 2010).

In application to SAIPE data for county 5-17 poverty, the regression variables used in $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ of eq. (9) include the constant 1 for an intercept term, and the following:

- logit of the proportion of child exemptions "in poverty" for the county, i.e., the logit of the ratio of the number of child exemptions claimed on tax returns whose adjusted gross income falls below the poverty threshold divided by the total number of child exemptions for the county;
- logit of an adjusted version of the county "tax child filer rate," which is defined as the number of child exemptions in the county claimed on tax returns divided by the county population age 0-17;
- logit of the ratio of county SNAP benefits recipients in July of the previous year to the county population of the previous year.

For $\mathbf{x}_{2i}$, which is used in the model for the (adjusted) ACS 5-year poverty estimates $\tilde{y}_{2i}$, we use the above variables defined for the middle year of the 5-year interval. Bell et al. (2007) did some evaluations of log-rate Fay-Herriot models for SAIPE county data with models that used the log-rate regression variables analogous to the above three logit-rate regression variables. Very similar untransformed ratios (one change: replacing the filer rate with its complement, the nonfiler rate) are used as regression variables in the SAIPE state model (Fay and Train 1995).

An issue arises in that the child tax filer rate often exceeds 1 due to the number of child exemptions for counties often exceeding their age 0-17 populations. This is primarily due to the fact that many child exemptions (especially college students) are older than 17. Compensating by increasing the upper age on the child population in the denominator of the filer rate creates another problem since small to medium size counties with major universities have a bulge in their age 18-24 populations from the resident college students. These students are not reflected in the university county's child exemption totals, but rather are spread around the counties of residence of the students' parents. Thus, replacing the 0-17 population with, say, the 0-24 population depresses the child filer ratios in these counties in a way that is unrelated to the counties' levels of poverty. Instead, we addressed the issue with child filer rates exceeding 1 by multiplying all county child filer rates by .54 prior to

the logistic transformation. This brought the highest of the child filer rates for 2011 to a number below 1, allowing the logistic transformation. We obtained similar results using untransformed child filer rates since, after the rescaling, the great majority of the child filer rates fell in the interval (.25,.75), a region where the logistic transformation is nearly linear.

Notice that we do not include the logit of the Census 2000 county estimate of the 5-17 poverty rate in $\mathbf{x}_{1i}$ or $\mathbf{x}_{2i}$. This is because our model effectively replaces it with the prior ACS 5-year estimates that are used in determining $\tilde{y}_{2i}$. While we could also include the Census 2000 estimates in our model (most appropriately as a third dependent variable in a trivariate model), preliminary results obtained with analogous linear models at the state level (along the lines of Huang and Bell (2012)) suggested this would be of little benefit.

4. Application of the Bivariate Binomial/Logit Normal Model to ACS 2011 estimates

We applied the bivariate binomial/logit normal model jointly to the ACS 2011 1-year county estimates and the ACS 2006-2010 5-year county estimates. We then used SAS's NLMIXED procedure to fit the model and compared model predictions to corresponding unraked predictions from the SAIPE county production model. In the future, we are also interested in raking the results from the proposed model to ascertain if less raking is required than with the production model.

The GVFs for the sampling variances of either the ACS 2011 1-year county estimates or 2006-2010 5-year county estimates were constructed as follows. Let

$$\mathrm{R}_i = \sum_{j=1}^{n_i} w_{ij}^2 \Big/ (\sum_{j=1}^{n_i} w_{ij})^2 \ ,$$

where w_{ij} is the survey weight of household j in county i, and n_i is the number of responding households in the sample for county i, $i = 1, \ldots, m$. This estimates the inverse of the effective sample size due (only) to differential weights for county i —see Kish (1987). The GVF for the sampling variance of $\hat{p}_i$, the direct ACS estimate of the county i 5-17 poverty rate, is defined separately for each dataset (2011 ACS 1-year and 2006-2010 ACS 5-year) as:

$$E(s_i^2) = \mathrm{GVF}_i = \gamma_0 (p_i(1-p_i))^{\gamma_1} (\mathrm{R}_i)^{\gamma_2}. \tag{11}$$

Initial estimates for the p_i were computed as $\breve{p}_i = g(\mathbf{x}_i'\hat{\eta}) = \exp(\mathbf{x}_i'\hat{\eta})/[1 + \exp(\mathbf{x}_i'\hat{\eta})]$ where $\hat{\eta}$ solves the nonlinear least squares problem

$$\min \sum_{i=1}^{m} (\hat{p}_i - g(\mathbf{x}_i'\eta))^2.$$

Since the $\hat{p}_i$ are the direct ACS poverty rate estimates, they can be zero. Note that $\breve{p}_i$ cannot be zero.

Taking logs of (11) and using the direct ACS sampling variance estimates as estimates of the left hand side of (11) yields a linear model which can be fit by regression. (We did this in R using the glm function.) Here we used all counties that did not have zero poverty counts and that met a certain sample size threshold (of 25 households, see Maples, 2012). The direct sampling variance estimator has large biases for small sample sizes, which can distort the model fit; hence the use of the threshold. Plugging $\breve{p}_i$ and the least squares estimates of γ_0, γ_1, and γ_2 into (11) provided predicted values $\widehat{\mathrm{Var}}(\hat{p}_i)$ that were the GVF sampling variance estimates for all counties, including those with $\tilde{y}_i = 0$. We then used these sampling variances estimates in the model defined by equations (5) and (6),

Table 1: Regression Coefficients and Standard Errors, Binomial/Logit Normal Model

Parameter	Estimate	Standard Error	T-statistic
α_1	0.083	0.030	2.76
β_{11}	0.732	0.032	22.56
β_{12}	-0.094	0.067	-1.41
β_{13}	0.295	0.025	11.76
α_2	0.0882	0.017	5.096
β_{21}	0.785	0.018	43.42
β_{22}	-0.200	0.024	-8.25
β_{23}	0.222	0.012	18.16
σ_{11}	0.274	0.012	23.27
σ_{22}	0.183	0.005	34.48
σ_{12}	0.017	0.003	6.84
ρ	0.3360	0.048	7.050

as described in the previous section, both when fitting the model and to obtain predictions of the true 2011 county school-age child poverty rates, p_{1i}.

We applied an iterative approach where, given the $\widehat{\text{Var}}(\hat{p}_i)$, we updated the $\breve{p}_i$ and then updated the $\widehat{\text{Var}}(\hat{p}_i)$, etc. We subsequently found, however, that the first iteration of this process appeared to have attained convergence. Appendix I is a detailed, step-by-step explanation on how we performed this iterative approach. Note that, as in the Fay-Herriot model, we assumed that sampling variances are known. In reality, we have estimates of the sampling variances through fitted GVFs, and errors in these estimates would be expected to produce biases in the predictions of poverty rates and their MSEs.

Table 1 shows the ML parameter estimates obtained by fitting the bivariate binomial/logit normal model (5) and (6) to the ACS data using SAS's NLMIXED procedure. Note that the estimate of the correlation coefficient $\rho = \sigma_{12}/\sqrt{(\sigma_{11}\sigma_{22})}$ is positive. If ρ were zero, the bivariate model would reduce to two independent univariate models, and the data from the 5-year ACS estimates would not affect prediction of the true 1-year poverty rates, p_{1i}. The only coefficient that is not statistically significant is β_{12}, the regression coefficient of logit(p_{1i}) on the logit of the tax child filer rate.

Figure 1 compares the production model county predictions of the proportion of school-aged children in poverty with those from the bivariate binomial/logit normal model. Figure 2 compares the respective prediction intervals. The predictions from the bivariate binomial/logit normal model are broadly similar to those of the current production model, while the corresponding prediction intervals tend to be a little wider, except for those counties with the widest intervals from the production model. Further investigation and comparisons to results from other alternative models discussed next will be pursued in future research.

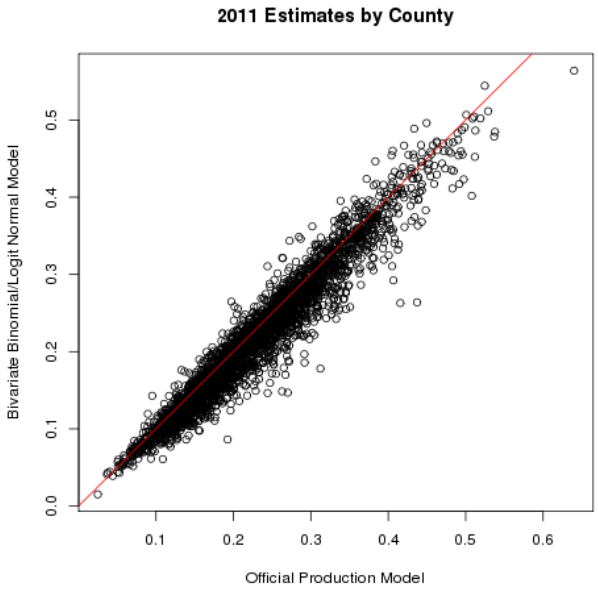

Figure 1: Comparison of Bivariate Estimates with Estimates from Production Model by County, 2011

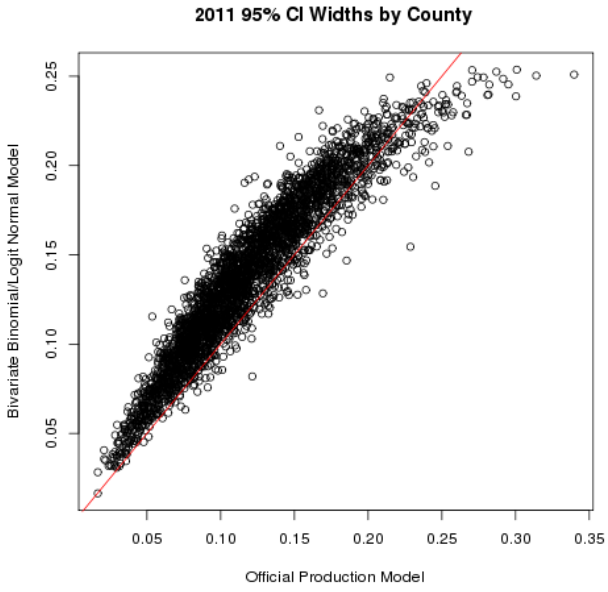

Figure 2: Comparison of Confidence Interval Widths by County, 2011

5. Future Research

In future research we plan to compare results from the bivariate binomial/logit normal model with results from the following additional models.

Log rate model: We consider a bivariate version of the linear Fay-Herriot model given by (1)–(2) where y_{1i} is the log of the ACS estimated 5-17 poverty rate for county i, and y_{2i} is the log of the prior ACS 5-year estimate of the 5-17 poverty rate for county i. Regression

variables in $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ include intercepts and the logs of the same ratios for which logits were taken for the bivariate binomial/logit normal model as described in Section 3. Sampling variances v_{1i} and v_{2i} for y_{1i} and y_{2i} are estimated directly by replication methods. For this model, it is necessary to drop from the model fitting the data for counties whose estimated 5-17 poverty rates are zero. As noted earlier, univariate versions of such models were investigated by Bell, et al. (2007).

Unmatched sampling and linking models (You and Rao 2002): Replace the Binomial assumption in eq. (8) with an assumption of normality. Letting $\hat{p}_{1i}$ and $\hat{p}_{2i}$ denote the ACS current 1-year and prior 5-year direct estimates of poverty rates, this becomes

$$\hat{p}_{1i}|p_{1i}, v_{1i} \sim N(p_{1i}, v_{1i}) \qquad \hat{p}_{2i}|p_{2i}, v_{2i} \sim N(p_{2i}, v_{2i}). \tag{12}$$

Nonlinear regression in the Fay-Herriot model: The model defined by eqs. (12) and (9) has random effects in the models for the logistically transformed population poverty rates. A simpler version keeps eq. (12), but replaces eq. (9) with Gaussian models for (untransformed) p_{1i} and p_{2i}:

$$p_{1i} = \frac{\exp(\mathbf{x}'_{1i}\beta_1)}{1+\exp(\mathbf{x}'_{1i}\beta_1)} + u_{1i} \qquad p_{2i} = \frac{\exp(\mathbf{x}'_{2i}\beta_2)}{1+\exp(\mathbf{x}'_{2i}\beta_2)} + u_{2i}$$

with the same assumptions made on $[u_{1i}, u_{2i}]'$ as for eq. (9). This model is unrealistic in that it allows for values of p_{1i} and p_{2i} that fall outside the interval $[0, 1]$. In the univariate version of this model, predictions of p_{1i} will be a weighted average of $\hat{p}_{1i}$ and the nonlinear regression prediction, and so must fall within $[0, 1]$. This is not true for predictions from the bivariate model, though how likely predictions are to fall outside $[0, 1]$ is unclear. Prediction intervals from either the univariate or bivariate models can certainly range outside $[0, 1]$.

Alternative link functions in the Bivariate GLMM model: This alternative modifies just eq. (9) by substituting a different link function for the logit. Common alternatives include the probit and the log-log (Agresti 1990).

Autoregressive Models Using Multiple Years of ACS Data: Instead of summarizing the information in 5 prior years of ACS data through the resulting 5-year estimates, a logical alternative to consider is to use the corresponding 5 individual 1-year estimates. Putting this together with the current 1-year estimates, this implies modeling 6 years of ACS 1-year estimates. We do this by extending (8)–(9) using a first-order autoregressive structure (AR(1)) as follows:

$$\tilde{y}_{it}|p_{it}, \tilde{n}_i \sim \text{Bin}(\tilde{n}_{it}, p_{it}) \qquad i = 1, \ldots, m, \quad t = 1, \ldots, T \tag{13}$$

$$\text{logit}(p_{it}) = \mathbf{x}'_{it}\beta_t + u_{it} \tag{14}$$

$$u_{it} = \phi u_{i,t-1} + \epsilon_{it} \tag{15}$$

where the ϵ_{it} are assumed $i.i.d.$ $N(0, \sigma^2_\epsilon)$, and $-1 < \phi < 1$ for stationarity. For simplicity, we label the years as $t = 1, \ldots, T$; our interest lies in predicting poverty rates in the final year, p_{iT}. Here we use $T = 6$ years of data, though one could obviously use more or fewer years.[2] With the stationarity assumption, the covariance matrix of $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})$ has

[2]Currently, we could use at most seven years of ACS data, since the ACS estimates start in 2005 and the last estimates available as of this writing are for 2011.

the general form (Box and Jenkins 1970, pp. 56-58)

$$\text{var}(\mathbf{u}_i) = \frac{\sigma_\epsilon^2}{1-\phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \phi \\ \phi^{T-1} & \phi^{T-2} & \cdots & \phi & 1 \end{bmatrix}.$$

Prediction MSEs from this model can be compared against those from the bivariate binomial/logit normal model (8)–(9) to assess the value of using the individual 1-year estimates. Analogous linear Gaussian models with AR(1) structure were investigated by Taciak and Basel (2012) for application to logs of ACS county 5-17 poverty estimates, and by Hawala and Lahiri (2012) for application to ACS estimates of county 5-17 poverty rates. Neither compared prediction MSEs against those of bivariate models applied to current ACS 1-year and prior ACS 5-year estimates, however.

6. APPENDIX: Steps for Implementation of the Bivariate Binomial/Logit Normal Model

Steps 1–7 below are applied separately for each of the model's two equations (for ACS 2011 1-year estimates and ACS 2006-2010 5-year estimates), so in these steps we omit from the notation subscripts to indicate the model equation. Steps 8 and 9 relate directly to the bivariate model.

1. For each county, compute
$$\mathrm{R}_i = \frac{\sum_{j=1}^{n_i} w_{ij}^2}{(\sum_{j=1}^{n_i} w_{ij})^2},$$
where w_{ij} is the weight of household j in county i, and n_i is the sample size of county i, $i = 1, \ldots, m$.

2. Compute $\hat{\eta}$ as the nonlinear least squares estimator through the following optimization problem:
$$min \sum_{i=1}^{m} (\hat{p}_i - g(\mathbf{x}_i'\eta))^2$$
where $\hat{p}_i$ $i = 1, \ldots, m$ are the direct county level 5-17 poverty rate estimates, and $g : \mathbb{R} \to (0, 1)$. Initially, g will be the inverse logit function.

3. Compute preliminary estimates $\breve{p}_i = g(\mathbf{x}_i'\hat{\eta})$, $i = 1, \ldots, m$ and note that, by construction, these cannot be outside the interval $(0, 1)$. Here $\mathbf{x}_i$ are the covariates used in our rate model.

4. Drop all counties that have less than 25 households in the sample. The number 25 is based on results from Maples (2012).

5. Using the counties meeting the threshold identified in Step 4 that do not have zero poverty counts, find an estimate of the sampling variance $\widehat{\mathrm{Var}}(\hat{p}_i)$ through the following model:
$$E(s_i^2) = \mathrm{GVF}_i = \gamma_0((p_i(1 - p_i))^{\gamma_1}(\mathrm{R}_i)^{\gamma_2}. \tag{16}$$
Taking logs this becomes
$$\log \mathrm{GVF}_i = \gamma_0^* + \gamma_1 \log(\breve{p}_i(1 - \breve{p}_i)) + \gamma_2 \log(\mathrm{R}_i)$$
which can be fit by linear regression.

6. Plug the $\hat{\gamma}_i$ obtained in step 5 into eq. (16) to produce GVF estimates $\hat{s}_i^2$ of s_i^2 for every county.

7. For each county, compute the effective sample size $\tilde{n}_i$ and the effective count of school aged children in poverty $\tilde{y}_i$ as
$$\tilde{n}_i = \breve{p}_i(1 - \breve{p}_i)/\widehat{\mathrm{Var}}(\hat{p}_i), \tag{17}$$
$$\tilde{y}_i = \tilde{n}_i \times \hat{p}_i. \tag{18}$$

8. Fit the bivariate binomial/logit normal model using the $(\tilde{n}_{1i}, \tilde{y}_{1i})$ and $(\tilde{n}_{2i}, \tilde{y}_{2i})$ from the previous step, with these values rounded to the nearest integers, and obtain new estimates for the p_{i1} and the p_{i2}.

9. Update the $\breve{p}_{1i}$ and $\breve{p}_{2i}$ using the results of Step 8. Repeat steps 5-8 until prediction results for p_{1i} and p_{2i} appear to have converged.

REFERENCES

Agresti, Alan (1990), *Categorical Data Analysis*, New York: John Wiley and Sons.

Bell, William R. (2000), "Models for County and State Poverty Estimates," Appendix A in *Small Area Estimates of School Age Children in Poverty: Evaluation of Current Methodology*, eds. Constance F. Citro and Graham Kalton, Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Washington, D.C.: National Academy Press, 169-184.

Box, George E.P. and Jenkins, Gwilym M. (1970), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

Bell, William R., Basel, Wesley, Cruse, Craig, Dalzell, Lucinda, Maples, Jerry, O'Hara, Brett, and Powers, David (2007), "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties," unpublished technical paper available at http://www.census.gov/did/www/saipe/publications/files/report.pdf.

Fay, Robert E. and Herriot, Roger 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.

Ghosh, Malay; Natarajan, Kannan; Stroud, T. W. F.; and Carlin, Bradley P. (1998), "Generalized linear models for small-area estimation", *Journal of the American Statistical Association*, **93** , 273-282.

Ghosh, Malay, Nangia, Narinder, and Kim, Dal Ho (1996), "Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach," *Journal of the American Statistical Association*, **91**, 1423-1431.

Fay, Robert E. and Train, George F. (1995) "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *Proceedings of the American Statistical Association, Government Statistics Section*, 154-159.

Hawala, Sam and Lahiri, Partha (2012), "Hierarchical Bayes Estimation of Poverty Rates," *Proceedings of the American Statistical Association, Survey Research Methods Section*, available at http://www.census.gov/did/www/saipe/publications/files/hawalalahirishpl2012.pdf.

Huang, Elizabeth T. and Bell, William R. (2012), "An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in SAIPE Models for Poverty Estimates," Research Report Number RRS2012-4, Center for Statistical Research and Methodology, U.S. Census Bureau, available at http://www.census.gov/srd/papers/pdf/rrs2012-04.pdf.

Kish, L. (1987). Weightng in Deft^2. *The Survey Statistician*. June, 1987.

Liu, B., Lahiri, P., and Kalton, G. (2007), Hierarchical Bayes Modeling of SurveyWeighted Small Area Proportions,*Proceedings of the American Statistical Association*, Survey Research Section, pp 3181-3186.

Maples, J., (2012) "An Examination of the Relative Variance of Replicate Weight Variance Estimators for Ratios Through First-Order Expansions", 2012 Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.

Maples, Jerry J. and Bell, William R. (2007), "Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data," Research Report Number RRS2007-11, Statistical Research Division, U.S. Census Bureau, available at http://www.census.gov/srd/papers/pdf/rrs2007-11.pdf.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

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

Otto, Mark C. and Bell, William R. (1995), "Sampling Error Modelling of Poverty and Income Statistics for States," *American Statistical Association, Proceedings of the Section on Government Statistics*, 160-165.

Plummer, Martyn (2010), "JAGS - Just Another Gibbs Sampler, JAGS 2.1.0.," May 12, 2010, available at https://sourceforge.net/projects/mcmc-jags.

Rao, J.N.K. (2003), *Small Area Estimation*, Hoboken, New Jersey: John Wiley.

SAS Institute Inc. (2010), "The NLMIXED Procedure," SAS version 9.2 Help and Documentation.

Slud, E.V. (2000) Models for Simulation and Comparison of SAIPE Analyses SAIPE Technical Report, http://www.census.gov/did/www/saipe/publications/techpubs.html

Slud, E. V. (2004) Small Area Estimation Errors in SAIPE using GLM versus FH Models. *Proc. ASA Section on Survey Research Methods* pp. 4402-4409.

Taciak, Jasen and Basel, Wesley (2012), "Time Series Cross Sectional Approach for Small Area Poverty Models," *Proceedings of the American Statistical Association, Survey Research Methods Section*, available at http://www.census.gov/did/www/saipe/publications/files/JTaciak%20Baseljsm2012.pdf.

You, Yong (2008), "An Integrated Modeling Approach to Unemployment Rate Estimation for Sub-Provincial Areas of Canada," *Survey Methodology*, **34**, 19-27.

You, Yong and Rao, J. N. K. (2002), "Small Area Estimation Using Unmatched Sampling and Linking Models," *The Canadian Journal of Statistics*, **30**, 3-15.