Small Area Estimation of Health Insurance Coverage in 2008 and 2009

Mark Bauder, Donald Luery and Sam Szelepka
Small Area Methods Branch
Social, Economic, and Housing Statistics Division
U.S. Census Bureau

February 16, 2012

1 Introduction

The Small Area Health Insurance Estimates (SAHIE) program at the U.S. Census Bureau produces estimates of numbers and proportions of those with and without health insurance coverage for demographic groups within states and counties. The demographic groups are defined by age, sex, and income, and in addition, for states by race and ethnicity. Income groups are defined in terms of income-to-poverty ratio (IPR), which is the family income divided by the appropriate Federal Poverty Level.

For 2008 and 2009, SAHIE publishes estimates for states for the following domains:

(1) The full cross classification of
   - 4 age categories: 0-64, 18-64, 40-64, 50-64
   - 4 race/ethnicity categories: all races, Hispanic, White not Hispanic, Black not Hispanic
   - 3 sex categories: all sexes, male, female
   - 5 income groups: all income, and IPR categories 0-138%, 0-200%, 0-250% and 0-400%.

(2) Age under 19 in IPR categories 0-138%, 0-200%, 0-250% and 0-400%.

For counties, SAHIE produces estimates for the same domains, except not for the 50-64 age category, and not by race/ethnicity.

The choice of domains is motivated by the needs of one of SAHIE's sponsors, the Centers for Disease Control and Prevention (CDC). The CDC has cancer screening programs for which the eligible population is low-income, uninsured women in specified age groups (SAHIE Team 2008). In addition, the age 0-19 low-income categories are relevant to the Children's Health Insurance Program (CHIP). Because the SAHIE models produce estimates for disjoint groups covering virtually everyone under age 65, we release estimates for men and women as well as children, and for other aggregates of possible interest.

The choice of the income groups 0-200% and 0-250% is motivated by the needs of the CDC and CHIP. The income groups 0-138% and 0-400% reflect the needs of recent heath care legislation. The Patient Protection and Affordable Care Act helps families gain access to health care by allowing Medicaid to cover families with incomes less than or equal to 138 percent of the federal poverty threshold. Also, families with incomes above the level needed to qualify for Medicaid, but less than or equal to 400 percent of the federal poverty threshold can receive tax credits that will help them pay for health coverage in the new health insurance exchanges.

In the sections to follow, we describe in detail the models used to produce the SAHIE estimates.

2 Overview of SAHIE modeling

2.1 The “base” level

We publish estimates for groups that sometimes overlap or are contained in one or another domain. However, actual modeling is done at a “base” level at which domains are disjoint, and are chosen so that the estimates needed for publication can be obtained as needed by aggregation. One exception to this general approach is that we do not attempt to make estimates for both age 0-17 and age 0-18 within the same model, by, for example, making estimates for age 0-17 and for age 18.

Setting aside the 0-18 age group for the moment, for states, we do the actual modeling for the full cross-classification of:

- 4 age categories: 0-17, 18-39, 40-49, 50-64
- 4 race/ethnicity categories: White not Hispanic, Black not Hispanic, Hispanic, and Other
- sex categories: male, female
- 5 income groups: 0-138% IPR, 138-200% IPR, 200-250% IPR and 250-400% IPR and above 400% IPR.

To obtain estimates for the 0-18 age group, we repeat the modeling procedure, replacing the 0-17 and 18-39 age groups with the 0-18 and 19-39 age groups. In the rest of this paper, we describe the modeling when the lowest two age groups are 0-17 and 18-39.

2.2 Two proportions to estimate

Let the acronym ARSH (age, race, sex, Hispanic origin) denote, for states, age by race/ethnicity by sex, and for counties, age by sex.

For states and counties, we have demographic population estimates for ARSH groups that we treat as known truth. To obtain estimates of the numbers with and without health insurance for states and counties within ARSH groups, we estimate two sets of proportions. Within each state or county by ARSH group, denoted by a, we estimate the proportions in each of the five income groups, $p_{ai}^{IPR}, i = 1, \ldots, 5$. In each ARSH by income group (a, i), we estimate the proportion insured, p_{ai}^{IC}. The number in ARSH by income group (a, i) is the product of p_{ai}^{IPR} and the population for ARSH group a. The number insured is the product of p_{ai}^{IC} and number from the ARSH/income group.

The SAHIE model consists of two largely distinct parts corresponding to these two proportions. We refer to the two parts of the model as the "income" and "insurance" parts.

2.3 Modeling survey data

The SAHIE model is an "area level" model (Rao 2003), in that it uses survey estimates for areas or domains of interest, rather than individual responses, and it uses other data that are aggregates rather than for individuals. Each of the two parts of the SAHIE model is similar to a well-known small area area-level model, the Fay-Herriot model. The Fay-Herriot model is a hierarchical model in which the variables of interest occupy a "middle" level, between high-level parameters such as regression coefficients, and observed data. Let $\theta_i, i = 1 \ldots n$ be the variable of interest, and $\hat{\theta}_i$ be a survey estimate of θ_i. A simple version of the Fay-Herriot model can be written

$$\hat{\theta}_i = \theta_i + \varepsilon_i \tag{1}$$

$$\theta_i = x_i^T \beta + u_i \tag{2}$$

where $\varepsilon_i \stackrel{indep}{\sim} \mathcal{N}(0, v_i^S)$ and $u_i \stackrel{indep}{\sim} \mathcal{N}(0, v_i^M)$.

The ε_i are sampling errors with sampling variances v_i^S. The u_i can be viewed as model errors, or as area-specific random effects, with model variance v_i^M. All the ε_i's are independent of all the u_i's. The equation in (1) is referred to as the "sampling model" and the equation in (2) is the "linking model."

In a frequentist context, assuming the sampling and model variances are known, it can be shown that the best linear unbiased predictor (BLUP) of each θ_i is a weighted average of the survey estimate and the regression prediction

$$\hat{\theta}^{BLUP} = \gamma x_i^T \hat{\beta} + (1 - \gamma)\hat{\theta}_i \tag{3}$$

$$\text{with} \;\; \gamma = \frac{v_i^S}{v_i^S + v_i^M} \tag{4}$$

Here $\hat{\beta}$ is the usual weighted least squares estimate from the regression of the $\hat{\theta}_i$'s on the x_i's. Note that γ becomes closer to one as the model variance becomes smaller relative to the sampling variance so that the BLUP estimate of θ_i is primarily the regression prediction estimate $x_i^T\hat{\beta}$. Complementarily, γ becomes closer to 0 when the model variance becomes large relative to the sampling variance so that the BLUP estimate is primarily the survey estimate $\hat{\theta}_i$.

A similar result holds in the Bayesian context. Conditional on the β's and variances, the mean of each θ_i is a weighted average of $\hat{\theta}_i$ and a regression prediction $x_i^T\beta$, with the weight on $x_i^T\beta$ being $\dfrac{v_i^S}{v_i^S + v_i^M}$.

In addition to the fact that the SAHIE model contains two parts, each of which is similar to a Fay-Herriot model, there are several differences between the SAHIE model and a standard Fay-Herriot model:

- In both the income and insurance parts of the model, we model a survey estimate $\hat{p}$ of the proportion p, but we assume that the logit of p satisfies a normal linear model. The sampling model and the linking model are not "matched."

- The survey estimates of the proportions in the income groups within an ARSH group are not independent. There are five income groups within each ARSH group. The survey estimates for those five must add to one. We model four of them, and assume the correlations correspond to those of a multinomial distribution.

- In the insurance part of the model, we do not assume that the logit(p)'s are independent. We instead assume that they have a block diagonal variance matrix with identical blocks

- In the insurance part of the model, we do not assume that the survey estimates $\hat{p}$ are distributed as normal, but instead that they follow a mixture of discrete and continuous distributions.

- We model as random some auxiliary data rather than treat them as fixed predictors in a regression.

In later sections, we give details of the model, including details of the items above. In the next section, we give a fuller discussion of the last item above.

2.4 Modeling auxiliary data

One large difference between the SAHIE model and the standard Fay-Herriot model in (1) and (2) is in how some non-survey data are used. In small-area models such as the Fay-Herriot model, "auxiliary" (i.e, non-survey) data are typically used as covariates to help predict the variables of interest. In the standard Fay-Herriot model, these covariates occur as fixed predictors as in the x_i in (2).

Fisher and Gee (Fisher 2003 and Fisher and Gee 2004) proposed an alternative in the context of estimating poverty. In their research for the Small Area Income and Poverty Estimates program, they developed an alternative to the usual Fay-Herriot model, that they refer to as an "error-in-variables" model. In their model, they treat the covariates as measures of the quantity of interest, θ (log poverty in their example), that are possibly biased and have random error. Let i index the observations and $j = 1, \ldots p$ index the p auxiliary data. Their model is

$$\hat{\theta}_i = \theta_i + e_i \tag{5}$$

$$A_{ij} = b_j + c_j\theta_i + u_{ij} \quad j = 1, \ldots, p \tag{6}$$

$$\theta_i \sim \mathcal{N}(\mu, v^\theta) \tag{7}$$

where the e_i and u_{ij} are mean zero error terms that are normal with variances that possibly depend on parameters. In this approach, the auxiliary data A_{ij} are treated in a way very similar to survey estimates. The primary difference is that they are not unbiased measures of θ, and thus the model includes the unknown parameters b_j and c_j.

A feature of the model in (5) - (7) is that the influence of $\hat{\theta}_i$ and the A_{ij} on the estimate of θ_i can vary observation to observation, depending on the relative sizes of the variances. This is an extension of the property noted in (3) and (4) in which the influences of the survey estimate and a regression prediction vary depending on the relative magnitude of their variances.

The approach of Fisher and Gee was extended to small area estimates of insurance coverage in Fisher, O'Hara, and Riesz (2006). The SAHIE model includes both fixed predictors $\mathbf{x}_i$ of the quantities of interest and auxiliary data to be modeled, $\mathbf{A}_i = (A_{i1}, \ldots, A_{ip})^T$.

The $\mathbf{A}_i$ are possibly nonlinear regressions of the θ_i, and the θ_i are modeled by a generalized linear model. The two parts of the SAHIE model each have the form

$$\begin{aligned} \widehat{\theta_i} &= \theta_i + e_i \\ A_{ij} &= h_j(\theta_i) + u_{ij} \quad j = 1, \ldots, p \\ g(\theta_i) &= \mathbf{x}_i^T\beta + v_i\ . \end{aligned}$$

where the e_i, u_{ij}, and v_i are error terms terms with mean zero and variances that depend on parameters, and that are independent except for exceptions noted later.

3 The primary data

We use the following primary data sources for states and counties.

ACS direct estimates. We have two sets of direct estimates from the American Community Survey:

- estimates of the number in each of the five IPR categories by age by race/ethnicity by sex categories for states, and by age and sex categories for counties
- estimates of the proportions insured in ARSH by income categories.

Census 2000 estimates. We use Census 2000 sample estimates of the number in geography/ARSH/IPR categories.

Internal Revenue Service exemption data. We use the number of Internal Revenue Service (IRS) exemptions in age by IPR categories in each state and county. The age categories are 0-17, 18-64, and 65+. We do not have actual ages for the IRS data. We use the number of child exemptions as a proxy for age 0-17 or for 0-18.

Supplementary Nutrition Assistance Program data. For each state and county, we use counts of the number of people participating in the Supplemental Nutrition Assistance Program (SNAP, formerly Food Stamps) from the United States Department of Agriculture.

Medicaid/CHIP participation data. We use Medicaid participation records from the Centers for Medicare and Medicaid Services (CMS). States submit their data to the CMS quarterly. Individuals are in the file if Medicaid covered them for at least one day during the quarter. We have Children's Health Insurance Program (CHIP) participation counts from states and counties gathered from a web page of the Centers for Medicare and Medicaid Services (CMS). We combine the Medicaid and CHIP participation data, and use the combined data for each state and county in cross-classifications of age by sex.

Demographic population estimates. We use demographic estimates of the resident population from the U.S. Census Bureau's Population Estimates Program. These estimates are published for the nation, states, and counties by age, sex, race, and Hispanic origin.

See http://www.census.gov/did/www/sahie/methods/inputs/index.html for more information about these data sources.

4 Model details

In this section, we describe in detail the components of the SAHIE model for states. There are some differences in the modeling of counties that we describe later. We use the following notation:

- ARSH (age, race, sex, Hispanic origin) refers to either an age by race by sex (for states) or an age by sex (for counties) category.
- a indexes state or county by ARSH category.
- i indexes IPR category.
- S_j denotes the sample size for the j^{th} category.
- POP denotes a demographic population estimate.
- N denotes a number of people. N_{ai}^{IPR} denotes the number of people in the a^{th} state or county by ARSH and i^{th} IPR category, and N_{ai}^{IC} denotes the number of people with health insurance coverage in the a^{th} state or county by ARSH and i^{th} IPR category. $N_{ai}^{UI} \equiv N_{ai}^{IPR} - N_{ai}^{IC}$ is the number uninsured.
- $p_{ai}^{IPR} \equiv N_{ai}^{IPR}/POP_a$ is the proportion among those in the a^{th} state or county by ARSH group who are in the i^{th} IPR category.
- $p_{ai}^{IC} \equiv N_{ai}^{IC}/N_{ai}^{IPR}$ is the proportion among those in the state or county by ARSH by IPR category (a, i) who have health insurance coverage.
- β denotes regression parameters that appear in a model for the proportion in an IPR category or a model for the proportion insured.
- α denotes a mean parameter, i.e., a parameter that appears in a model for the mean of the Census 2000 or administrative record data.
- λ denotes a variance parameter, i.e., a parameter that appears in the model for the sampling variance of the ACS estimates, or in a model for the variance of the Census 2000 or administrative record data.
- Hatted variables such as $\hat{p}_{ai}^{IPR}$ denote direct survey estimates.
- Overlines such as $\overline{CEN}$ denote means.

The parameters α and λ typically depend on one or more of the age, race/ethnicity, sex, or IPR categories. We suppress indices that show these dependencies.

The income part of the model allows us to estimate p_{ai}^{IPR}, the proportion of people in IPR category i, within state or county by ARSH category a. The number of people within the state or county by ARSH by income group is given by $N_{ai}^{IPR} = p_{ai}^{IPR} POP_a$ The insurance part of the model allows us to estimate p_{ai}^{IC}, the proportion insured within state or county by ARSH by IPR category (a, i). We combine these to estimate the primary quantities of interest, N_{ai}^{IC} and N_{ai}^{UI}, the number insured and the number uninsured, where

$$N_{ai}^{IC} = p_{ai}^{IC} N_{ai}^{IPR}$$
$$N_{ai}^{UI} = (1 - p_{ai}^{IC}) N_{ai}^{IPR}$$

4.1 The income part of the state model

4.1.1 Modeling ACS estimates of proportions in income groups

In the first part of the income model, we model $\hat{p}_{ai}^{IPR}$, the ACS estimate of the proportion in IPR category i within state by ARSH category a. We assume that these ACS estimates are unbiased and normally distributed. Note that for any a, $\sum_{i=1}^{5} \hat{p}_{ai}^{IPR} = 1$. For this reason, we model four of the five IPR categories, and do not treat those four as independent. We assume a parametric model for the variances, and assume the correlations correspond to those of a multinomial distribution. The following is the model for the ACS estimates of proportions in IPR categories.

$$(\hat{p}_{a1}^{IPR}, \ldots, \hat{p}_{a4}^{IPR})' \,|\, p^{IPR}, \lambda \sim \mathcal{N}((p_{a1}^{IPR}, \ldots, p_{a4}^{IPR}),\ \Sigma_a^{IPR})$$

$$\left(\Sigma_a^{IPR}\right)_{ii} = \frac{\lambda_1 p_{ai}^{IPR}(1 - p_{ai}^{IPR})}{S_a^{\lambda_2}}\ ,\quad i = 1, \ldots, 4$$

$$\left(\Sigma_a^{IPR}\right)_{ij} = \rho_{aij} \sqrt{\left(\Sigma_a^{IPR}\right)_{ii} \left(\Sigma_a^{IPR}\right)_{jj}}$$

$$\text{where } \rho_{aij} = -\sqrt{\frac{p_{ai}^{IPR} p_{aj}^{IPR}}{(1 - p_{ai}^{IPR})(1 - p_{aj}^{IPR})}} \qquad i \neq j$$

The parameters λ_0 and λ_1 vary by age.

4.1.2 The regression part of the income model

We assume that p_{ai}^{IPR}, the proportion of those in state by ARSH group a who are in IPR group i, follows a five-category logistic model with normal errors. Let X^{IPR} be a matrix of fixed predictors with rows $(x_{ai}^{IPR})^T$.

$$p_{ai}^{IPR} = \frac{\exp(\mu_{ai}^{IPR})}{\sum_{i=1}^{5} \exp(\mu_{ai}^{IPR})}$$

$$\mu_{ai}^{IPR} \,|\, \beta^{IPR}, v^{M,IPR} \overset{indep}{\sim} \mathcal{N}\left((x_{ai}^{IPR})^T \beta^{IPR}, v^{M,IPR}\right)$$

$$\text{with } (x_{a5}^{IPR})^T = 0 \text{ for all } a.$$

The model variance $v^{M,IPR}$ is the same for all a and i.

The predictors in X^{IPR} for states are

- main effects for IPR
- two-way interactions between age and IPR, between race/ethnicity and IPR, and between sex and IPR

- three-way interactions among age, race/ethnicity, and IPR
- three-way interactions among age, sex, and IPR
- three-way interactions between age, IPR and the following continuous variables:
  - the logit of the proportion who are Hispanic in the state (from demographic population estimates)
  - the median log IPR, as measured by tax records
  - the variance of log IPR, as measured by tax records.

4.1.3 Modeling state Census 2000 estimates, IRS exemptions, and SNAP counts

We model the means of the Census 2000 estimates, the IRS exemptions, and the SNAP counts as functions of N_{ai}^{IPR}, the number of people in IPR category i, within state by ARSH, a.

4.1.4 Modeling Census 2000 estimates for states

We model the Census 2000 estimates, CEN_{ai}^{IPR}, of the number of people in state by ARSH by IPR categories. We assume these estimates have means, $\overline{CEN}_{ai}$, that are linear functions of the N_{ai}^{IPR}, and are conditionally independent.

$$
\begin{aligned}
CEN_{ai} | \, N^{IPR}, \alpha, \lambda &\overset{indep}{\sim} \mathcal{N}(\overline{CEN}_{ai}, v_{ai}), \text{ where} \\
\overline{CEN}_{ai} &= (\alpha_0 + \alpha_1) N_{ai}^{IPR} \\
v_{ai} &= \lambda_0 \overline{CEN}_{ai}^{\lambda_1}
\end{aligned}
$$

The α's and λ's are parameters to be estimated. The parameter α_0 varies by age and IPR, while α_1 varies by race and IPR. The variance parameter λ_0 varies by age by race/ethnicity, and λ_1 takes on one value.

4.1.5 Modeling IRS exemptions for states

From the IRS, we have the number of IRS exemptions by state by two approximate age categories (0-17, 18-64) by IPR categories. The age categories are approximate because the number that we use for the age 0-17 category is actually the number of child exemptions.

We assume that the numbers of exemptions are normally distributed with a mean that is a linear function of aggregate N_{ai}^{IPR}'s, and are conditionally independent. Let t index state by the two age categories.

$$
\begin{aligned}
TAX_{ti} | \, N^{IPR}, \alpha, \lambda_0, \lambda_1 &\overset{indep}{\sim} \mathcal{N}\left(\overline{TAX}_{ti}, \; v_{ti}\right) \\
\overline{TAX}_{ti} &= \alpha N_{ti}^{IPR} \\
v_{ti} &= \lambda_0 \overline{TAX}_{ti}^{\lambda_1}
\end{aligned}
$$

where N_{ti}^{IPR} is the number of people in state by age by IPR category ti. N_{ti}^{IPR} is obtained by summing N_{ai}^{IPR} over the appropriate age, race/ethnicity, and sex categories. The parameters α and λ_0 vary by the two ages by IPR, and λ_1 has one value.

4.1.6 Modeling SNAP participation for states

SN_s is the number of SNAP participants by state. We model the mean, $\overline{SN}_s$, as a linear function of the number of people in the state in the IPR 0-138% category. We assume that the SN_s's are normally distributed and conditionally independent. Let s index state. Then

$$\begin{aligned} SN_s | \, N^{IPR}, \alpha, \lambda &\overset{indep}{\sim} \mathcal{N}\left(\overline{SN}_s, v_s\right) \\ \overline{SN}_s &= \alpha N_{s1}^{IPR} \\ v_s &= \lambda_0 \overline{SN}_s^{\lambda_1}. \end{aligned}$$

Here N_{s1}^{IPR} is the number of people in a state in the 0 - 138% IPR category. The parameters α, λ_0, and λ_1 each take one value.

4.2 The insurance part of the state model

In the insurance part of the model, we model ACS estimates of p_{ai}^{IC}, the proportion insured in the state by ARSH by IPR category, and the combined Medicaid/CHIP data. From this part of the model, we obtain estimates of p_{ai}^{IC}, which enables us to estimate our primary quantities of interest, N_{ai}^{IC} and N_{ai}^{UI}, the number insured and the number uninsured in state by age by race/ethnicity by sex by IPR category ai, by $N_{ai}^{IC} = p_{ai}^{IC} N_{ai}^{IPR}$ and $N_{ai}^{UI} = (1 - p_{ai}^{IC}) N_{ai}^{IPR}$.

4.2.1 Modeling the ACS estimates of the proportion insured

Proportions insured are often close to one. ACS estimates of proportions insured are often one, sometimes zero, and are bounded between zero and one. Rather than assume normality, we model the ACS estimates of proportions insured in a way to capture that they are bounded, have positive probability mass at zero and one. We use the term "three-part model" for the model we use, following Pfeffermann et al. (2008) who use the term "two-part model" to refer to a similar model.

We model the probability that $\hat{p}_{ai}^{IC}$ is one, the probability that $\hat{p}_{ai}^{IC}$ is zero, and conditional on $0 < \hat{p}_{ai}^{IC} < 1$, we assume that $\hat{p}_{ai}^{IC}$ follows a beta distribution. Let $p_{ai}^{(0)}$ and $p_{ai}^{(1)}$ be the probabilities that $\hat{p}_{ai}^{IC}$ is zero and one. The model is

$$\hat{p}_{ai}^{IC} | \, p^{IC}, \lambda, \zeta \begin{cases} = 0 & \text{with probability } p_{ai}^{(0)} \\ = 1 & \text{with probability } p_{ai}^{(1)} \\ \sim \text{Be}(a_{ai}, b_{ai}) & \text{with probability } 1 - p_{ai}^{(0)} - p_{ai}^{(1)} \end{cases} \tag{8}$$

with

$$\text{var}(\hat{p}_{ai}^{IC}) = \frac{\lambda_1 p_{ai}^{IC}(1 - p_{ai}^{IC})}{S_{ai}^{\lambda_2}} \tag{9}$$

$$p_{ai}^{(0)} = (1 - p_{ai}^{IC})^{1+\zeta_0(S_{ai}-1)} \tag{10}$$

$$p_{ai}^{(1)} = (p_{ai}^{IC})^{1+\zeta_1(S_{ai}-1)} \tag{11}$$

where Be denotes the beta distribution. The parameters λ_1 and λ_2 vary by age and IPR, as do the parameters ζ_0 and ζ_1. Note that the parameters, a_{ai} and b_{ai}, of the beta distribution in (8) are functions of p_{ai}^{IC}, $p_{ai}^{(0)}$, $p_{ai}^{(1)}$ and $\text{var}\left(\hat{p}_{ai}^{IC}\right)$. We chose the functions for the variance and the probabilities of zero and one in (9) - (11) by starting with what the variances and probabilities of zero and one, would be under simple random sampling, and introducing parameters to accommodate the effects of non-independence due to the sample design and correlated responses. Szelepka and Bauder (2011) considered various groups of observations and compared within groups the predicted and actual frequencies of survey estimates of zero and one. They found close agreement, confirming the choice of functions in (10) and (11).

4.2.2 The regression part of the insurance model

The model for the proportions insured is logistic-normal with a multivariate error structure. Let $\mu_a^{IC} = (\mu_{a1}^{IC}, \ldots, \mu_{a4}^{IC})^T$, and let $X^{(a)}$ be the matrix made up of the five rows $(x_{a1}^{IC})^T, \ldots, (x_{a5}^{IC})^T$, and X^{IC} the data matrix obtained by stacking the $X^{(a)}$'s.

$$\begin{aligned} p_{ai}^{IC} &= \text{logit}^{-1}\left(\mu_{ai}^{IC}\right) \\ \mu_a^{IC} \mid \beta, \Sigma^{IC} &\overset{indep}{\sim} \mathcal{N}(X^{(a)}\beta, \Sigma^{IC}) \end{aligned}$$

where Σ^{IC} is a 5×5 matrix whose elements are estimated. The predictors in X^{IC} for states are

- main effects for age, race/ethnicity, sex and IPR
- all two-way interactions among age, race/ethnicity, sex and IPR
- three way interactions among age, sex, and IPR
- the state median log IPR, as measured by tax data, interacted with IPR and with two ages (children, adult) by IPR
- the variance within a state of log IPR, as measured by tax data, interacted with IPR and with two ages (children, adult) by IPR

4.2.3 Modeling Medicaid/CHIP enrollees

Let MED_m be the number of people enrolled in Medicaid or CHIP in a state by age by sex category, denoted m. We assume that the mean, $\overline{MED}_m$, is a function of the number insured in IPR 0-250%. We assume that the Medicaid counts MED_m are independent, conditional on all N^{IC}_{ai} and parameters. We have

$$\begin{aligned} MED_m|\, \gamma, \alpha, \lambda &\sim \mathcal{N}\left(\overline{MED}_m, v_m\right) \\ \overline{MED}_m &= \gamma_s \alpha N^{IC}_{m1} \\ \gamma_s|\, \delta &\sim \text{Gamma}\,(\text{mean} = 1, \text{var} = \delta) \\ v_m &= \lambda_0 \overline{MED}_m^{\lambda_1} \end{aligned}$$

where s is the state of the m^{th} observation. N^{IC}_{m1} is obtained by summing N^{IC}_{i1} over the race/ethnicity categories and the IPR categories 0-138%, 138-200% and 200-250%. The parameter α varies by age by sex, λ_0 varies by age, and λ_1 takes one value. The γ_s's are state level random effects with variance, δ, and are independent given δ. The γ_s's are multiplicative, rather than additive, effects to ensure that the coefficients of N^{IC}_{m1} are always positive, while still allowing the possibility that the γ_s's reduce the coefficient on N^{IC}_{m1}.

5 The county model

For counties, the models for the ACS estimates of proportions in IPR categories and of proportions insured are like those in sections 4.1.1 and 4.2.1. The model for Medicaid/CHIP participation is like that in 4.2.3.

The regressions in the income and insurance parts of the model for counties have different predictors than for states.

5.1 Predictors for county IPR and IC regressions

The predictor matrix X^{IPR} for counties (as in section 4.1.2 for states) includes the following:

- main effects for IPR
- two-way interactions between age and IPR, and between sex and IPR
- the three-way interactions among age, sex, IPR
- log county population interacted with IPR, and with age by IPR (the coefficients can differ for small and large counties)
- logit of the proportion Hispanic based on demographic estimates, interacted with IPR and with age by IPR
- county median log IPR, interacted with IPR and with age by IPR
- variance of log IPR, interacted with IPR and with age by IPR
- state, interacted with IPR.

Define age2 to take two values: one for age 0-17, the other for any of the other age groups: 18-38, 40-49, 59-64. The predictor matrix X^{IC} for counties (as in section 4.2.2 for states) includes the following.

- IPR, age and sex categories and all their two- and three- way interactions
- state interacted with age2
- each of the following county level variables, and its interactions with age, IPR, and age2 by IPR:
  - log county population
  - county median log IPR
  - variance of log IPR
  - logit of the proportion of adults with less than high school education
  - logit of the Census 2000 estimate of the proportion who are non-citizens
  - logit of the Census 2000 estimate of the proportion who are American Indian/Alaskan Native
  - logit of the proportion of adults who are employed by firms of size 19 or less
  - logit of the proportion of adults who are employed by firms of size 100 or more

5.2 Modeling county Census 2000 estimates, IRS exemptions, and SNAP counts

As with states, we model the means of the Census 2000 estimates, the IRS exemptions, and the SNAP counts as functions of the N_{ai}^{IPR}, summed to the appropriate level. However, there are notable differences in how we model these data for counties.

5.2.1 Modeling the Census 2000 estimates for counties

For counties, we model the Census estimates, CEN_{ai}, of numbers in county by ARSH group a and IPR category i as follows.

$$\begin{aligned} CEN_{ai} | \, \alpha, \lambda &\sim \mathcal{N}(\overline{CEN}_{ai}, v_{ai}) \\ \overline{CEN}_{ai} &= \alpha_0 \left(N_{ai}^{IPR}\right)^{\alpha_1} \\ v_{ai} &= \lambda_0 \overline{CEN}_{ai}^{\lambda_1} \end{aligned}$$

where CEN_{ai} is the Census estimate. The parameter α_0 varies by age and IPR, α_1 varies by age, λ_0 vary by age and IPR, and λ_1 varies by age.

5.2.2 Modeling IRS exemptions for counties

Let t index county by the three tax age categories. For counties, we have

$$\begin{aligned} TAX_{ti} | \, \nu, \alpha, \lambda &\sim T\left(\nu, \text{mean} = \overline{TAX}_{ti}, \text{var} = v_{ti}\right) \\ \overline{TAX}_{ti} &= \alpha_0 N_{ti}^{IPR} \\ v_{ti} &= \lambda_0 \overline{TAX}_{ti}^{\lambda_1} \end{aligned}$$

where T is the t-distribution, parameterized in terms of the degree of freedom parameter, ν, and the mean and variance. N_{ti}^{IPR} is obtained by summing N_{ai}^{IPR} over the appropriate age and sex categories. We use a t-distribution here because when we fit the model assuming normality, some residuals were too large to be consistent with the normality assumption. We did not observe this with states. The parameters α, λ_0 and ν vary by the three age by IPR categories. There is one value for λ_1.

5.2.3 Modeling SNAP participation for counties

Let c index county. For SNAP data, we have

$$\begin{aligned} SN_c \mid N^{IPR}, \alpha, \lambda &\sim \mathcal{N}\left(\overline{SN}_c, v_c\right) \\ \overline{SN}_c &= \alpha_0 \left(N_{c1}^{IPR}\right)^{\alpha_1} \\ v_c &= \lambda_0 \overline{SN}_c^{\lambda_1}. \end{aligned}$$

Note that as with states, we predict SNAP participation from only the lowest IPR category. The parameters $\alpha_0, \alpha_1, \lambda_0$, and λ_1 each take one value.

5.3 Prior distributions

For the Bayesian modeling, we generally use vague priors for the high level parameters. For the regression coefficients β^{IPR} and β^{IC}, we use the (improper) uniform prior over the real numbers of appropriate dimension. For multiplicative parameters in functions for means, we use truncated normal distributions with large variances. For multiplicative variance parameters (λ_0 in most cases above) we use the (improper) prior $\frac{1}{\sqrt{\lambda_0}}$. For parameters that are exponents in variance functions, we use a uniform prior on $(0, 3)$.

6 Model selection

We made many modeling decisions to arrive at the current SAHIE models. In addition to the overall form of the model, these decisions include choices of predictors, mean and variance functions, and distributions. We describe some of the criteria we used in the next sections.

6.1 Model diagnostics

6.1.1 Standardized residuals

Some choices of mean, variance, and density functions resulted from perceived lack of fit based on diagnostics we use. Our primary model diagnostic is a certain type of standardized residual. For the survey estimates, Census and administrative data that we model, we predict means and variances so that for any data, y, that we model, we can obtain a form of standardized residual and squared residual

$$E_{\theta|data}\left[\frac{y - E(y|\,\theta)}{\sqrt{\mathrm{var}(y|\,\theta)}}\right] \quad \text{and} \quad E_{\theta|data}\left[\frac{(y - E(y|\,\theta))^2}{\mathrm{var}(y|\,\theta)}\right] \tag{12}$$

from the Markov chain Monte Carlo (MCMC) output used to fit the model. See Chib and Greenberg (1995) for an explanation of MCMC. If the model is correct and y is normally distributed, this standardized residual is distributed as approximately normal(0,1). The standardized squared residuals should have a mean of approximately one. We check that averages of these residuals over large groups of observations are close to zero, and check for extremely small or large values. We look at plots against various quantities such as the predicted mean, population, predicted variance, and where appropriate, sample size. We also look at boxplots of standardized residuals for different values of categorical variables such as age and IPR, and against quantiles of population. We check that the averages over large groups of squared standardized residuals are reasonably close to one.

6.1.2 Posterior predictive p-values

Another model diagnostic that we use is the posterior predictive p-value (PPP-value) (Gelman, Meng, and Stern (1996)). A posterior predictive p-value is a measure of how surprising or improbable some function of the data (and possibly parameters) is, under the posterior predictive distribution of that data. Let y represent all of the data and θ represent all of the parameters. A PPP-value is defined as $P_{y^{rep},\theta|y}(T(y^{rep},\theta) \geq T(y,\theta))$ for some function T where the probability is with respect to $p(y^{rep}|\theta)p(\theta|y)$, the joint distribution of a replication of the data, y^{rep}, and θ, conditional on y. Let y_i represent a single data point. We use the functions $T_1(y,\theta) = y_i$ and $T_2(y) = (y_i - E(y_i|\theta))^2$. Thus, the PPP-value corresponding to T_1 is $P_{y^{rep},\theta|y}(y_i^{rep} \geq y_i)$. We refer to this PPP-value as the PPP-value for the mean because many values near 0 or near 1 suggest that means given by the model are generally too low, or too high, respectively. We refer to the PPP-value corresponding to T_2 as the PPP-value for the variance since it measures the surprise in the squared distance between the data and its mean. We compute PPP-values for each of the data sources in the model. We look at plots of PPP-values against various quantities, such as population, posterior means, posterior variances, and sample sizes. Our approach is to use the PPP-values informally to check for evidence of model failure. Many values near zero or near one would suggest problems with the model.

6.2 Selecting predictors for the regression parts of the income and insurance models

In order to select predictors for the income and insurance parts of the model, we generally consider the posterior means and variances of the regression coefficients. We form an approximate 95% credible interval for the regression coefficient by taking its posterior mean plus or minus two times its posterior standard deviation. Generally speaking, we include a predictor in the model if the approximate 95% credible interval does not include zero.

7 Benchmarking

We benchmark SAHIE estimates of the numbers insured and uninsured in order to make them consistent with a set of national ACS estimates, and to make county estimates consistent with state estimates. We benchmark state estimates to a relatively small set of national direct estimates of numbers insured and uninsured. We benchmark all possible county estimates to the corresponding state estimates. The benchmarking procedure for counties is a simple proportional adjustment. The procedure for states is more complex.

7.1 State to national benchmarking.

We benchmark the state estimates to ACS national estimates of insured and uninsured for the following categories:

- IPR 0-250%
- age 0-17, IPR 0-250% (or age 0-18, IPR 0-200%)
- age 18-64 (or age 19-64)
- Hispanic
- not Hispanic
- White not Hispanic
- Black not Hispanic.

7.1.1 Methodology for state to national benchmarking

The benchmarking procedure we use was developed by Luery (1986) in the context of controlling survey weights to control totals. The procedure is as follows. Let B be the number of benchmarks (here, 14), and let $\hat{\mathbf{N}} = (\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_B)'$, be the benchmarks. Let S be the number of small area, or model, estimates, and let $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S)'$ be those estimates. We want to adjust the model estimates so that their sums over states equal the benchmarks. Let b index the benchmarks, let i index the area (here, state by ARSH by IPR by insured/uninsured). Let $\mathbf{X} = (x_{ib})$

be the S x B matrix such that $x_{ib} = 1$ when area i contributes to benchmark b, and 0 otherwise. Then the adjusted estimates $\hat{Y}_i^*$ meet the constraints when $\sum_{i=1}^{S} x_{ib}\hat{Y}_i^* = \hat{N}_b$ for all b.

We want a set of benchmarked estimates that are, in some sense, optimal. Generally, benchmarked estimates are preferable when they are close to the original estimates. We choose to minimize the relative quadratic loss function

$$\sum_{i=1}^{S}(\hat{Y}_i^* - \hat{Y}_i)^2/\hat{Y}_i \;. \tag{13}$$

That is, we minimize the squared change from the original to the benchmarked estimate, relative to the size of the original estimate. It can be shown that there exists a unique set of $\hat{Y}_i^*$ that sum to the benchmarks and minimize (13). This optimal set of benchmarked estimates, $\hat{\mathbf{Y}}^* = (\hat{Y}_1^*, \hat{Y}_2^*, \ldots, \hat{Y}^*)'$ is given by

$$\hat{\mathbf{Y}}^* = \hat{\mathbf{Y}} + \mathbf{D}(\hat{\mathbf{Y}})\mathbf{X}\mathbf{P}(\hat{\mathbf{N}} - \mathbf{X}^{\mathbf{T}}\hat{\mathbf{Y}}) \tag{14}$$

where $\mathbf{D}(\hat{\mathbf{Y}})$ is a diagonal matrix with the entries of $\hat{\mathbf{Y}}$ along the diagonal and $\mathbf{P} = [\mathbf{X}^{\mathbf{T}}\mathbf{D}(\hat{\mathbf{Y}})\mathbf{X}]^{-\mathbf{1}}$.

For the i^{th} area, this can be written as

$$\hat{Y}_i^* = \hat{Y}_i(1 + \sum_{b=1}^{B} f_b x_{ib}) \tag{15}$$

where the f_b are the B factors given by $\mathbf{F} = (f_b) = \mathbf{P}(\hat{\mathbf{N}} - \mathbf{X}^{\mathbf{T}}\hat{\mathbf{Y}})$. Thus, the choice of the relative quadratic loss function ensures that if two areas i and i' have the same indicators, that is, if $x_{ib} = x_{i'b}$ for all b, then they receive the same proportional change to their estimates, as given in (15).

7.2 Methodology for county to state benchmarking

We benchmark county estimates so that in each state, the county estimates for insured and uninsured in each age by sex by IPR group sum to the benchmarked state estimates. For each cross-classification of age, sex, and income, we apply an adjustment factor to the county estimates of the number insured and the number uninsured so that the sum of the county estimates equals the state estimate. Let c index counties, j index age by sex categories, i index income categories, and s index states. The adjusted estimate of the numbers insured and uninsured are given by

$$\hat{N}_{cji}^{IC,adjusted} = \frac{\hat{N}_{sji}^{IC}}{\sum_c \hat{N}_{cji}^{IC}}\hat{N}_{cji}^{IC} \qquad \hat{N}_{cji}^{UI,adjusted} = \frac{\hat{N}_{sji}^{UI}}{\sum_c \hat{N}_{cji}^{UI}}\hat{N}_{cji}^{UI}$$

where $\hat{N}_{sji}^{IC}$ and $\hat{N}_{sji}^{UI}$ are state estimates of the insured and uninsured for age by sex by income categories,and the sums are over the counties, c, in state s.

References

Bauder, D.M. and Szelepka, S. (2011), "A Three-Part Model for Survey Estimates of Proportions", to appear in *2011 American Statistical Association Proceedings of the Section on Survey Research Methods*.

Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, 49, 327-335.

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.

Fisher, R. (2003), "Errors-In-Variables Model for County Level Poverty Estimation", SAIPE Working Paper, Washington, DC, U.S. Census Bureau.
http://www.census.gov/did/www/saipe/publications/files/tech.report.5.pdf

Fisher, R. and Asher, J. (2000), "Alternate CPS Sampling Variance Structures for Constrained and Unconstrained County Models", SAIPE Technical Report #1, Washington, DC, U.S. Census Bureau.
http://www.census.gov/did/www/saipe/publications/files/tech.report.1.revised.pdf

Fisher, R. and Gee, G. (2004), "Errors-In-Variables County Poverty and Income Models", *2004 American Statistical Association Proceedings of the Section on Government and Social Statistics*.
http://www.census.gov/did/www/saipe/publications/files/
FisherGee2004asa.pdf

Fisher, R., O'Hara, B. and Riesz, S. (2006), "Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups", *2006 American Statistical Association Proceedings of the Section on Government and Social Statistics*.

Fisher, R. and Turner, J. (2003), "Health Insurance Estimates for Counties", *2003 American Statistical Association Proceedings of the Section on Survey Research Methods*.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733-807.

Luery, D. (1986), "Weighting Survey Data Under Linear Constraints on the Weights", *1986 American Statistical Association Proceedings of the Section on Survey Research*

Methods, 325-330.

Pfeffermann, D., Terryn, B., and Moura, F. (2008). Small Area Estimation under a Two-Part Random Effects Model with Application to Estimation of Literacy in Developing Countries.*Survey Methodology*, 34, 233-247.

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

Small Area Health Insurance Estimates Team, U.S. Census Bureau (2008), "The Feasibility of Publishing County-level Estimates of the Number of Women Eligible for the CDCs NBCCEDP",
http://www.census.gov/did/www/sahie/publications/files/
cdc_feasibility_report_oct2008.pdf.