

# SAHIE 2006 State Demographic and Income Model Methodology

## Overview

The documentation presented here will be expanded upon in a forthcoming technical paper. Technical papers that describe previous versions of the model are available on the Publications page.

We estimate the number of people with health insurance coverage by state within demographic and income groups, and estimate the number without insurance as the difference between estimates of the number of people in a group and the number with insurance in that group. The number insured in a group is the product of the number in the group and the proportion in that group who are insured. Correspondingly, our model has two main parts: one for estimating the number of people in state demographic and income groups, and one for estimating the proportion with health insurance in these groups. Each part is a hierarchical two-level regression model. We use Bayesian methods to estimate the parameters in the model. Our estimates take into account that the estimates from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) for different states have different reliabilities due to varying sample sizes in each state. Our estimates for states with large sample sizes tend to be close to the CPS ASEC estimates. The demographic and income groups of the CPS ASEC estimates that are modeled are described in the *Model Details* section.

The dependent variables in the regression models are:

- 3-year average estimates from the CPS ASEC.
  - numbers of people in demographic and income groups.
  - proportions insured in these groups.
- Estimates from Census 2000 – Sample Data (i.e., the long form).
- Numbers of IRS tax exemptions.
- Numbers of Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamp program, participants .
- Numbers of Medicaid and Children's Health Insurance Program (CHIP) participants.

The CPS ASEC estimates of the number of people in a state demographic and income group, and of the proportion insured in that group, are assumed to be unbiased. The other dependent variables are related to these numbers or proportions but are not assumed to be unbiased estimates of them. However, they are believed to be predictive of them. For more information about the dependent variables, see information about data inputs on the web page.

The universe for these health insurance estimates is the CPS ASEC poverty universe. Therefore, we use demographic population estimates (from the U.S. Census Bureau's Population Estimates Program) that are adjusted to the CPS ASEC poverty universe to ensure that the small area health insurance estimates conform to this universe. For more information about the demographic population estimates, see information about data inputs.

We provide on the website margins of error for the estimates that represent the uncertainties associated with both sampling and modeling. These margins of errors can be used to construct 90-percent confidence intervals which are approximate Bayesian credible regions calculated using posterior standard deviations and a normal approximation.

### **Model Details**

We estimate the number of people with health insurance coverage by state, age, race/ethnicity, sex, and income groups. The income groups are defined by the ratio of family income to the Federal Poverty Level (FPL), which is referred to as the income-to-poverty ratio (IPR). The number of people in a group with health insurance can be factored as the product of the number of people in the group and the proportion of these people who have health insurance. Correspondingly, there are two submodels: one for estimating the number of people in the groups and another for estimating the proportion of people in these groups who have health insurance. The number of people without health insurance is estimated as the difference between the estimates of the number of people in the group and the number with health insurance.

We provide estimates for age groups 0 to 64, 18 to 64, 40 to 64, and 50 to 64 and for 0 to 18. The demographic and IPR groups of the CPS ASEC estimates that are modeled for the first four age groups are:

- Age: 0-17, 18-39, 40-49, 50-64, 65+ (the 65+ group is used only in the model for the number of people by demographic and IPR groups).
- Race/ethnicity: White non-Hispanic, Black non-Hispanic, Hispanic, Other non-Hispanic.
- Sex: Female, Male.
- IPR: family  $IPR \leq 200\%$ ,  $200\% < IPR \leq 250\%$ ,  $IPR > 250\%$ .

The IPR groups above are defined so that a single model can be used to publish estimates for  $IPR \leq 200\%$  and  $IPR \leq 250\%$ . The development of these estimates is funded in part by the Centers for Disease Control and Prevention's (CDC's) National Breast and Cervical Cancer Early Detection Program (NBCCEDP). Based on the CDC's programmatic requirements, we publish estimates for both of these IPR groups.

For the estimates of the 0 to 18 age group, the demographic and IPR groups are:

- Age: 0-18, 19-39, 40-49, 50-64, 65+ (the 65+ group is used only in the model for the number of people by demographic and IPR groups).
- Race/ethnicity: White non-Hispanic, Black non-Hispanic, Hispanic, Other non-Hispanic.
- Sex: Female, Male.
- IPR: family  $IPR \leq 200\%$ ,  $200\% < IPR \leq 250\%$ ,  $IPR > 250\%$ .

A separate run was done to provide estimates for this age group.

*To simplify the discussion, the following methodology will be described only in terms of the age groups 0-17, 18-39, 40-49, 50-64, 65+.*

## **Model for the Number of People for States by Demographic and IPR Groups**

The model is a multivariate, two-level hierarchical model. The data sources are:

- Three-year average estimates of the numbers of people from the CPS ASEC by state and demographic and IPR group.
- The estimates from Census 2000 – Sample Data for the same groups as the CPS ASEC.
- The number of tax exemptions for nine groups in each state defined by age groups 0-17, 18-64, and 65+ crossed with  $IPR \leq 200\%$ ,  $200\% < IPR \leq 250\%$ , and  $IPR > 250\%$  groups. (For simplicity, we say the age groups are 0-17, 18-64, and 65+; actually, they are defined less exactly on tax forms. For more information about the tax data, see information about data inputs.)
- The number of SNAP participants by state.

### **First Level**

At the first level, we model the data sources conditional on the numbers of people in the state demographic and IPR groups. These numbers are latent variables – unknown quantities – that are to be estimated by the model. Each data source is modeled as a regression where the independent variables are these latent variables. In all of the models, the errors are modeled as normally distributed and are assumed to be independent across age, race/ethnicity, sex, and IPR groups, as well as between the data sources.

- The CPS ASEC estimate in a group is modeled such that its expected value is the numbers of people in that group and its variance is the sampling variance. We assume that the sampling variance has a particular functional form, which contains parameters that need to be estimated.
- The 2000 Census – Sample Data estimate in a demographic and IPR group is modeled as proportional to the number of people in that group. There are different proportionality factors for different age, race/ethnicity, and IPR groups. The combined model and census sample data variance is modeled as proportional to the expected number of people raised to a power.
- The tax exemption data are broken down into nine age and income groups for each state. Each age and income group is modeled as a linear regression where the independent variables are the number of people in the demographic and IPR groups comprising the tax exemption group. The model variance is modeled as proportional to the expected number of people in the tax exemption group raised to a power.
- The SNAP participants in a state are modeled as proportional to the number of people in a state in families with  $IPR \leq 200\%$  raised to a power. The model variance is modeled as proportional to the expected number of SNAP participants raised to a power.

## Second Level

At the second level, we model the number in states by demographic and IPR groups. We do this by modeling the proportion of those in a state by demographic group who are in each of the IPR groups. These proportions are then multiplied by the demographic population estimates to obtain the estimates for the state by demographic and IPR groups.

The proportion of people in the state by demographic and IPR group is modeled as a three-category logistic regression model with independent normal errors. We assume the errors have constant variance. The independent variables for the model are:

- Sex, age, and race/ethnicity each crossed with IPR.
- Three-way effects of sex, age, and IPR; and age, race/ethnicity and IPR.
- Interactions of tax non-filing rates in state and age groups with IPR (the tax non-filing rate is computed as the difference between the demographic population estimate and the number of tax exemptions divided by the demographic population estimate).
- Interactions of the proportion Hispanic in the state with IPR.
- Interactions of the demographic population estimate for the state with IPR.

## Model for the Proportions of People with Health Insurance Coverage for States by Demographic and IPR Groups

The model for the proportion with health insurance coverage is a two-level hierarchical model. The data sources are:

- The CPS ASEC estimates of the proportion of people with health insurance for state by demographic and IPR groups.
- The number of Medicaid and Children's Health Insurance Program (CHIP) participants in each state, for groups defined by age and sex.

## First Level

At the first level, the CPS ASEC estimates and the number of participants in Medicaid and CHIP are modeled, conditional on the proportion insured. The proportion insured is a latent variable – unknown quantity – that is to be estimated by the model.

The CPS ASEC estimated proportion with health insurance is modeled such that the expected value of the CPS ASEC proportion is the proportion of people with health insurance in each of the corresponding groups. The errors are assumed to be normally distributed and independent. The variances are the sampling variances. We assume that the sampling variance has a particular functional form, which contains parameters that need to be estimated.

The Medicaid and CHIP participation data are broken down into eight age by sex groups for each state. The number of participants in each age by sex group is modeled as proportional to the number of people with health insurance coverage and  $IPR \leq 200\%$  within the age by sex group.

The proportionality constants are products of fixed and random effects. The errors are assumed to be independent of each other, and independent of the CPS ASEC estimates. The model variance is modeled as proportional to the expected number of Medicaid and CHIP participants in the state age by sex group raised to a power.

## **Second Level**

The proportion of people with health insurance in state by demographic and IPR groups is modeled as a logistic regression model with independent normal errors. We assume the errors have constant variance. The independent variables for the model are:

- An intercept term.
- Main effects for age, race/ethnicity, sex, and IPR.
- Main effect of each state.
- Two-way interactions of age, race/ethnicity, and sex with IPR groups.
- Two-way interactions of age with sex and race/ethnicity with sex.

## **Prior Distributions**

All high-level parameters, such as regression coefficients and variance parameters are given prior distributions. The prior distributions for the regression coefficients are noninformative flat priors. The prior distributions for the other parameters are normal or gamma priors that generally carry little information.

## **Model Limitations**

A model is an approximate, not exact, description of the distribution of the data. The models have been evaluated against the data and no major discrepancies have been found between the predictions from the model and the data. Research continues to improve the models so that they more accurately describe the distributions of the data. For example, modeling choices, including assumptions of independence, the choices of variance forms, estimation of the sampling variances, and the use of two variables derived from tax data, have not been completely validated. Because the models are determined using the same data used to produce the estimates, and because the model used is one of many possible models for the data, we may underestimate variances of the estimates.

## **Estimation of the Number of People with Health Insurance Coverage**

We use a Bayesian approach for estimating the parameters in the model and the number of people with health insurance. The estimated number of people with health insurance is the posterior mean conditional on the CPS ASEC data, the Census 2000– Sample Data, the tax exemption data, the SNAP data, and the Medicaid and CHIP participation data. The final estimate for a state demographic and IPR group is a complex mixture of the CPS ASEC estimate for that group and the other data. Estimates with large sample sizes, and thus low variances, tend to be close to the CPS ASEC estimates.

The method used to estimate the parameters in the model and to estimate the number of people with and without health insurance is called Markov Chain Monte Carlo (MCMC). This method involves drawing samples from the posterior distribution of the parameters in the model and the posterior distribution of the number of people with and without health insurance. Estimates for the number of people are the averages from the samples, called the posterior means. In order to control these estimates so that they agree with national CPS ASEC estimates, this procedure was altered as described in the section *Controlling to National CPS ASEC Estimates*.

### **Controlling to National CPS ASEC Estimates**

We control the estimates from the SAHIE State models to national CPS ASEC estimates of the number of people with health insurance, and to national CPS ASEC estimates of the number of people without health insurance, for the following demographic groups:

- 0-64, Hispanic
- 0-64, non-Hispanic
- 0-64, White non-Hispanic
- 0-64, Black non-Hispanic
- 0-64, IPR  $\leq$  250%
- 18-64
- 0-17, IPR  $\leq$  250%

(In order to provide estimates for ages 0-18 and IPR  $\leq$  200%, we changed the last two controls to 19-64 and 0-18, IPR  $\leq$  200%.) The choice of the above controls has the effect of also controlling to the groups 0-64; 0-17; 0-64, other non-Hispanic; 0-64, IPR  $>$  250%; and 0-17, IPR  $>$  200%.

For example, the estimate for insured 0-64, Hispanics can be obtained by summing the small area estimates for people with health insurance over all states, all age groups less than 65, the Hispanic origin group, both sexes, and all IPR groups.

We control to national estimates for two reasons. One is to guarantee consistency with CPS ASEC estimates at national levels of aggregation. The second reason is to control for possible weaknesses or failures of the model. We account for the variances of the controls by treating the controls as random quantities in the estimation program.

### **Measure of Errors and Confidence Intervals**

One goal of the small area work is to provide measures of uncertainty surrounding the estimates. We provide on the website margins of error for the estimates that represent the uncertainties associated with both sampling and modeling. These margins of errors can be used to construct 90-percent confidence intervals which are approximate Bayesian credible regions calculated using posterior standard deviations and a normal approximation.