SAHIE 2001 Demographic and Income Model Methodology

Overview

We estimate the number of people with health insurance coverage by state within demographic groups and income categories. 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 numbers of people in state demographic and income groups, and one for estimating the proportions with health insurance in these groups. Each part is a hierarchical two-level regression model. We use Bayesian methods to estimate the model. We estimate the number without insurance as the difference between the number of people in a category and the number with insurance. The demographic groups and income categories are described in the Model Details section.

The dependent variables in the regression models are:

- 3-year average estimates from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS);
- estimates from Census 2000;
- numbers of IRS tax exemptions;
- numbers of participants of the Supplemental Nutrition Assistance Program (SNAP) formerly known as the Food Stamp program; and
- 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, are assumed to be unbiased. The other dependent variables are related to and indicative of these numbers or proportions but are not assumed to be unbiased estimates for them.

The universe for these health insurance estimates is the CPS poverty universe. Therefore, we use demographic estimates of the population adjusted to the CPS poverty universe.

For further information on the dependent variables and population estimates, see information about data inputs.

We control the estimates for states so the following conditions are met:

- The numbers of insured sum to national CPS ASEC direct estimates of insured in the demographic groups. As a result, the numbers of uninsured sum to the national CPS ASEC direct estimates of uninsured.
- The sum of the insured and uninsured over all income categories for a state demographic group equals the CPS poverty universe for that state demographic group.

The CPS ASEC estimates for different states have different reliability because of the size of samples in each state. Our estimates consider this factor. Estimates from states with larger samples tend to be closer to the direct estimates.

We provide a confidence interval for each estimate that represents uncertainty from both sampling and modeling. These confidence intervals are Bayesian credible regions calculated using posterior standard deviations and a normal approximation.

Model Details

We estimate the number of people by state, age, sex, race/ethnicity, and income categories. The income groups are defined by the ratio of family income to the Federal Poverty Level (FPL), which is referred to as income-to-poverty ratio (IPR). To make the necessary estimates, the model estimates the number of people with health insurance in each state by demographic categories for specified IPR categories. The demographic categories are defined by age, race/ethnicity, and sex.

The number of people in a category with health insurance can be factored as the product of the number of people in the category and the proportion of these people who have health insurance. Correspondingly, there are two submodels: one for estimating the number of people in the categories and another for estimating the proportion of people in these categories who have health insurance. The number of people without health insurance is estimated as the difference between the number of people in the category and the number with health insurance.

For modeling, the demographic groups and IPR categories are defined as follows:

- Age: 0-17, 18-39, 40-49, 50-64;
- Race/ethnicity: White non-Hispanic, Black non-Hispanic, Hispanic, Other;
- Sex: Female, Male; and
- IPR: family IPR $\leq$ 200%, 200% < IPR $\leq$ 250%, IPR > 250%.

The age groups are defined so that estimates for ages 18-64, 40-64, and 50-64 can be made. The IPR groups are defined so that a single model can be used to make estimates for IPR $\leq$ 200% and IPR $\leq$250% for each state. Development of these estimates is partly funded by the Centers for Disease Control and Prevention's (CDC's) Division of Cancer Prevention and Control (DCPC) for the National Breast and Cervical Cancer Early Detection Program (NBCCEDP). At this point we are only publishing estimates for women.

For previous publications describing initial and intermediate results see the demographic research section of the publications page.

Model for the Number of People for States by Demographic Group and IPR Categories

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

- the estimated numbers of people from the CPS ASEC by state and demographic group and IPR category;
- the estimates from Census 2000 long form for the same categories as the CPS ASEC;

- the numbers of tax exemptions for four groups in each state defined by age groups 0-17 and 18+ crossed with IPR ≤ 200% and IPR > 200% categories; and
- the number of Supplemental Nutrition Assistance Program (SNAP) participants, formerly known as the Food Stamp program, by state.

First Level

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

- The CPS ASEC estimate in a category is modeled such that its expected value is the number of people in that category and the variance is the modeled sampling variance.
- The Census 2000 estimate in a category is modeled to be proportional to the number of people in that category. There are different proportionality factors for different race/ethnicity groups. The model and census sampling variance are jointly modeled as proportional to the number of people.
- The tax exemption data is broken down into four age and income categories for each state. Each age and income category is modeled as a linear regression where the independent variables are the number of people in a subset of the demographic and IPR categories comprising the tax exemption category. The model variance is modeled as proportional to a power of 1.7 of the number of people in the state and age group. The factor of 1.7 was chosen because it fit well in tests of the model.
- The SNAP participants in a state are modeled as proportional to the number of people in a state in families with IPR ≤ 200%. The model variance is modeled as proportional to the number of SNAP participants.

Second Level

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

The proportions of people in the state by demographic group who are in IPR categories are modeled as a multiple-category logistic regression with independent normal errors. The independent variables for the model are:

- sex, age, and race/ethnicity each crossed with the IPR categories;
- three-way effects of sex, age, and IPR; sex, race/ethnicity and IPR; and age, race/ethnicity and IPR; and
- interactions of the state tax non-filing rates with the IPR categories.

The model variance is modeled as constant.

Model for the Proportions of People with Health Insurance Cross-Classified by State, Demographic Groups, and IPR Categories

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

- the estimated proportions of people with health insurance from the CPS ASEC by state and demographic and IPR categories; and
- the numbers of Medicaid and Children's Health Insurance Program (CHIP) participants in each state, for categories defined by age and sex.

First Level

At the first level, the CPS ASEC direct estimate and the number in Medicaid or CHIP are modeled, conditional on the proportions insured. They are each modeled as a regression model with the proportion insured as the independent variable. The errors are normally distributed and independent between age, sex, race and IPR categories, as well as between the direct estimate and Medicaid/CHIP. The proportions are latent variables - unknown quantities - that are to be estimated by the model.

The CPS ASEC estimated proportions with health insurance are modeled such that the expected values of the CPS ASEC proportions are the unknown proportions of people with health insurance in each of the corresponding categories and the variances are the sampling variances.

The Medicaid and CHIP participation data is broken down into eight age by sex categories for each state. Each age by sex category is modeled as a linear regression where the independent variables are the numbers of people with insurance coverage and IPR $\leq$ 200% for a subset of the demographic categories contained within the age by sex category. The model variance is modeled as proportional to the number of people in the state and Medicaid age by sex groups.

Second Level

The proportion of people with health insurance in state by demographic group and IPR category is modeled as a logistic regression with normal errors. The independent variables for the model are:

- one-way effects of age, race/ethnicity, sex, and IPR categories;
- two-way interactions of age, race/ethnicity, and sex with IPR categories;
- two-way interactions of age by sex and race by sex; and
- state indicators.

The model variance is modeled as constant.

Prior Distributions and Model Limitations

Prior Distributions

All high level parameters are given prior distributions. The prior distributions for the regression coefficients are noninformative flat priors. The prior distributions for the other parameters are gamma priors that generally carry little information.

Model Limitations

For both models, modeling choices, including assumptions of independence, the choice of variance forms, and the use of two variables - tax exemptions and non-filing rate - derived from tax data, have not been completely validated. Hence, we may have underestimated variances of the estimates.

Estimation of the Number of People with Health Insurance

We use a Bayesian approach for estimating the models and estimating the number of people with health insurance. The estimated numbers of people are the posterior means conditioned on the CPS ASEC data, the Census 2000 data, the tax exemption data, the food stamp data, and the Medicaid/SCHIP data. The final estimate for a state demographic group by IPR category is a complex mixture of the direct CPS ASEC estimate for that group and the other data. Estimates with large sample sizes or very high insurance rates, and thus low variances, tend to be closer to the direct estimates.

Controlling to the National CPS ASEC Estimates

The last step in the production process is controlling the state estimates to the national CPS ASEC estimates for people with and without health insurance, within demographic groups. The numbers of insured are aggregated to the national level for each of the demographic groups, and the ratios of the national CPS ASEC direct estimates to the aggregated national model-based estimates are calculated to form the raking factors. A raking factor is multiplied with the health insurance estimates for the corresponding demographic group in every state and IPR group to form the raked health insurance estimates. The uninsured estimates are formed by subtracting these estimates from the estimated number of people in the corresponding state demographic by IPR categories.

We control to national estimates for two reasons. One is to guarantee consistency with direct CPS ASEC estimates at high 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.

Standard Errors and Confidence Intervals

One goal of the small area work is to provide measures of uncertainty surrounding the estimates. The model-based estimates shown in the tables are accompanied by their 90-percent confidence intervals constructed using a normal approximation from the estimated posterior standard deviations. The confidence intervals are Bayesian credible regions. Confidence interval half-widths for estimated numbers are rounded up to preserve coverage probabilities.