Abstract

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces income and poverty estimates using decennial census data, household survey data, administrative records and population estimates. This research focuses on augmenting the SAIPE program's county poverty model with an additional predictor, Medicaid participant data, from the Centers for Medicare and Medicaid Services. Since Medicaid eligibility is means tested based on income and assets, many Medicaid participants may live in families that have income below the federal poverty threshold. Medicaid participant totals might then have utility in estimating county poverty levels. Model diagnostics are presented under both the current SAIPE model and the expanded model for year 2000 to gauge possible model improvement.

Keywords: small-area estimation, small-domain areas, SAIPE, Medicaid, income, poverty

1. Introduction

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces estimates of the number of people in poverty at the school district, county and state levels. The SAIPE program’s poverty models employ both direct survey-based estimates of poverty levels from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) and regression predictions of poverty levels based on administrative records and Census 2000 data. This document is part of ongoing research to improve SAIPE program methodology.

A perennial challenge to small area estimation is obtaining data that are consistently available over time, closely related to the subject of interest and geographically relevant for the small area. In estimating poverty levels, the SAIPE program has in the past found federal tax information and Food Stamp Program data to meet these criteria. Recently, micro-level Medicaid participant data (stripped of personal identifiers) have become available to the Census Bureau from the Centers for Medicare and Medicaid Services (CMS). These Medicaid data may have utility in estimating county poverty levels, and this relationship is herein explored.

Section 2 below provides background regarding the Medicaid program. Section 3 discusses the data used in this work and examines correlations in the data. Section 4 presents experimental poverty models and results from estimation with the additional Medicaid variable. Section 5 concludes and discusses further modeling possibilities.

2. Medicaid Program

"Medicaid is a program that pays for medical assistance for certain individuals and families with low incomes and resources" (CMS, 2005b). The program is funded through state and federal sources, with many program expenses receiving a federal dollar match. States are required to provide Medicaid benefits to "categorically needy" groups, such as Temporary Assistance for Needy Families (TANF) recipients, Supplemental Security Income (SSI) recipients, most children under age 19 in families with income near or below the federal poverty threshold and several other groups. Criteria for the "categorically needy" vary little from state to state.

States may additionally provide Medicaid benefits to "categorically related" groups, such as infants and pregnant women with income up to 185% of the federal poverty threshold, certain institutionalized individuals, certain elderly, blind or disabled adults, the medically needy and many other groups. Benefits for many "categorically related" groups are not required to be extensive and can be "quite restrictive" (CMS, 2005b). Eligibility criteria for "categorically related" groups vary across states. The medically
needy provision is particularly broad in interpretation, and some states may participate in waiver programs that cover targeted groups (non-SSI disabled, HIV patients, etc.) whose income or assets might exceed the standard program limits.

In 1997, as part of the Balanced Budget Act, the State Children’s Health Insurance Program (SCHIP) was created to expand health insurance coverage among children. States receive federal funding in order to implement SCHIP either by expanding Medicaid eligibility for children or by setting up separate state children’s health insurance programs. States may also enact a combination of these two options (HHS, 2002). The criteria for SCHIP eligibility vary from state to state, but they are generally less stringent than those for the traditional Medicaid program.

3. Data

3.1 Medicaid Participation and Poverty Status

Data on Medicaid participants correlate positively with measured poverty levels perhaps due to income and asset restrictions that determine whether a person may participate in the Medicaid program. In applying for Medicaid benefits, many applicants attest to having income below the federal poverty threshold, and applicant income is subject to verification through the Income and Eligibility Verification System (CMS, 2004). Still, some people who are in poverty may not be eligible for the Medicaid program because there are other criteria than income alone. For instance, states may generally deny Medicaid coverage to healthy working-age adults without children regardless of poverty status.

Medicaid participants are people who are enrolled in the Medicaid program. Some people who might meet the eligibility requirements for the Medicaid program do not apply. Such would-be participants may not know about the program, may feel it is too difficult to access or understand, may be concerned about possible stigma or impact on other benefits or may be confident in their physical health (Ellwood, 1999). Such factors somewhat limit the Medicaid data’s correlation with measured poverty levels.

Although states receive federal dollar matching for many Medicaid program expenses, states have discretion in the provision of benefits, and this allows eligibility rules and covered medical procedures to vary by state (Coughlin and Zuckerman, 2002). States may also vary in their efforts for neighborhood outreach, which may lead to uneven Medicaid program awareness by the public (Schwalberg, et al., 1999). While Medicaid participant data may provide a broad indicator of county poverty levels, the relationship between individual Medicaid participation and poverty status is probably somewhat different in each state.

Figure 1 below is a scatter diagram of county Medicaid participant totals (average of April, May and June of 2000) and 3-year average CPS ASEC county poverty estimates centered on year 2000. The CPS ASEC estimates are for the 1,156 counties that have CPS ASEC sample and non-zero estimated numbers of people in poverty in at least one of the three years. More information about the CPS ASEC estimates is provided in Section 3.3. The Medicaid participant data are 3-month average county totals from April, May and June of 2000. These data are explained in greater detail in Section 3.2.

There is a strong positive correlation between Medicaid participant totals and CPS ASEC poverty estimates at the county level. Medicaid participant data might be used to help refine the SAIPE program’s county poverty models, and this possibility is investigated in Section 4.

3.2 Medicaid Data

The Balanced Budget Act of 1997 required states to report all Medicaid program eligibility and claims data to the Centers for Medicare and Medicaid Services (CMS), the administering agency for the Medicaid program, through the Medicaid Statistical Information System (MSIS) starting no later than
January 1, 1999 (CMS, 2005a). The Census Bureau has obtained micro-level Medicaid participant data (stripped of personal identifiers) from CMS based on these MSIS administrative records.

The MSIS data are received quarterly with three months of records in each file. The files are unedited and contain known data anomalies, many of which are documented on the CMS webpage, http://www.cms.hhs.gov/medicaid/msis/anomolies.pdf. Many of the anomalies are minor, and there is often no clearly suitable adjustment. Census Bureau quality control procedures look for duplicate records and delete those found, which in the past have been only a small share of the total number of records.¹

A variable designating SCHIP participation is included in the MSIS files. Some codes of this variable are denoted “optional,” and, therefore, the reporting of SCHIP participants in the MSIS files varies by county and state. Because of this varied reporting and because SCHIP participants generally face less stringent eligibility criteria than those in the traditional Medicaid program, I do not include SCHIP participants in the Medicaid tabulations I use.

The Medicaid data are stable from month to month in most counties. The results I present in Section 4.3 are not dependent upon the particular calendar month used for the Medicaid data. Leading or lagging the Medicaid data by up to a year does not materially impact the regression results. Facing a variety of possible monthly lags and averages, I use a three-month average of county Medicaid participant totals from April, May and June of 2000 as the Medicaid variable. My tallies count part-month eligibility as eligibility for the full month. If someone were eligible for one day of one of these three months, that person is counted fully for that one month, zeroed for the other two months, and computed as 0.333 participants.

The MSIS data offer a distinction between people who are eligible for the full scope of Medicaid benefits and those who are eligible for only a restricted subset of benefits (due to citizenship status, Medicare dual-eligibility, pregnancy-related services and other reasons). I create two separate Medicaid tallies: one that includes all Medicaid participants (Tally A) and one that includes only participants having full eligibility (Tally B).

### 3.3 Other Data

Descriptions of the data used in the SAIPE program’s county poverty models for year 2000 follow below.

**Poverty estimates from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS).** The CPS ASEC is an annual survey of nearly 100,000 households conducted each February, March and April that includes income questions about the prior calendar year. Documentation of the CPS ASEC is available at http://www.bls.census.gov/cps/.

The SAIPE program creates a 3-year weighted average of the CPS ASEC estimated number of people in poverty by county that is centered on year 2000 for year 2000 poverty estimates. This 3-year weighted average is computed as the product of the weighted 3-year average CPS ASEC poverty rate and the corresponding weighted 3-year average CPS ASEC poverty universe. The log of this 3-year average poverty estimate forms the dependent variable in the SAIPE program’s county poverty models.

Only counties with CPS ASEC sample and with non-zero estimated numbers of people in poverty in at least one of the three years can be used in fitting the regression model defined in Section 4.1. In 2000, of the 3,140 U.S. counties, 1,156 counties had CPS ASEC sample and non-zero estimated numbers of people in poverty in at least one of the three years.

County observations use weights that are adjusted so that all county observations are self-representing. This is achieved by multiplying the final CPS ASEC weights by the Primary Sampling Unit probability of selection. For details regarding self-representing counties and Primary Sampling Units, see Chapter 3 of “Technical Paper 63RV,” available at: http://www.census.gov/prod/2002pubs/tp63rv.pdf.

**Federal tax exemptions.** The total number of exemptions claimed on Internal Revenue Service (IRS) tax returns for income year 2000, including the filer, the spouse of the filer and any children or other dependents, is tallied for each U.S. county. Personal identifiers are stripped prior to data use.

**Federal tax exemptions on returns in poverty.** Poverty status is determined for each tax return by comparing the adjusted gross income to the appropriate family poverty threshold for 2000 (See http://www.census.gov/hhes/www/poverty.html). All exemptions per return are assigned the same poverty status.

¹ In the four quarters of data comprising calendar year 2000, no state had more than 50 duplicate records or greater than 0.02% duplication rate within a given quarter.
Total population estimates. County-level estimates of resident population as of July 1, 2001\textsuperscript{2} are obtained from the Census Bureau’s Population Estimates Program.

Food Stamp Program participants. County-level tallies of the number of people participating in the Food Stamp Program in July of 2000 are obtained from the Food and Nutrition Service in the U.S. Department of Agriculture.


4. Models and Estimates

4.1 Poverty Model

The SAIPE program’s county poverty model used for year 2000 estimates follows below in notation. A similar model is run separately for three age groups: related children ages 5-17 in families,\textsuperscript{3} children ages 0-17 and people all ages. In this document, only the “all ages” model is considered.

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \mu_i + \epsilon_i \]

\[ \mu_i \sim N(0, \sigma_{\mu_i}^2) \text{, independent across counties } i \]

\[ \epsilon_i \sim N(0, \sigma_\epsilon^2) \text{, independent across counties } i \]

Dependent Variable

\[ y_i = \log (3\text{-year weighted average of CPS ASEC estimated number of people in poverty in county } i) \]

Independent Variables

\[ x_{1i} = \log (\text{number of federal tax exemptions on returns in poverty in county } i) \]

\[ x_{2} = \log (\text{number of Food Stamp Program participants in county } i) \]

\[ x_{3i} = \log (\text{estimated total population in county } i) \]

\[ x_{4i} = \log (\text{number of federal tax exemptions in county } i) \]

\[ x_{5i} = \log (\text{Census 2000 estimated number of people in poverty in county } i) \]

Error

\[ \mu_i \text{ = sampling error for the dependent variable for county } i \text{ (assumed independent across counties)} \]

\[ \epsilon_i \text{ = model error for county } i \text{ (assumed independent across counties)} \]

The data vary numerically over a wide scale and are log-transformed, which makes their distributions more symmetric. Thus, the model is run in double-log format, where coefficient estimates represent the estimated percentage change in the dependent variable for a one-percentage change in an independent variable, holding the other independent variables fixed.

The SAIPE program includes variables \( x_3 \) and \( x_4 \) because the difference between \( x_3 \) and \( x_4 \) represents people not reported on IRS tax returns (i.e., people in non-filing families) (National Research Council, 2000). People with lower incomes might be less likely to file taxes (or be an exemption on a filer’s tax return) than people with higher incomes.

4.2 Model Estimation

The model is estimated with weighted least-squares regression and maximum likelihood estimation. The county weights are based on the sum of estimated sampling and model error variances. Observations from large counties (with many surveyed households), thus, receive more weight in the regression than do observations from smaller counties (with fewer surveyed households).

The sampling and model error variances underlying the county weights are estimated in several steps. First, an auxiliary model, similar in form to the main model above, is estimated, in which Census 2000 log poverty estimates form the dependent variable, Census 1990 log poverty estimates form an independent variable in place of Census 2000 log poverty estimates, and the remaining main-model independent variables form additional independent variables. Sampling error variances of the Census 2000 poverty estimates are estimated with a generalized variance formula.

\textsuperscript{2} The relevant population estimates for year 2000 poverty estimates are population estimates from the survey year, 2001, because CPS ASEC estimates are based on interviews conducted in February, March, and April of year 2001.

\textsuperscript{3} “Related children ages 5-17 in families” denotes children ages 5 to 17 who are in families in which these children are related to the householder by birth, marriage or adoption.
The auxiliary model is estimated using weighted least-squares regression in an iterative process that, starting with an entered value, produces sequentially better maximum-likelihood estimates of the beta coefficients and the model error variance until convergence. Then, two assumptions are made: 1) the model error variance in the main model is equal to the estimated model error variance from the auxiliary model and 2) the model error variance is constant across counties. The main model is then estimated through an iterative process that, upon convergence, provides maximum likelihood estimates of the beta coefficients and the total sampling error variance. The estimated total sampling error variance is apportioned to individual counties according the inverse of the square root of county CPS ASEC sample size.

Empirical Bayesian techniques are used to combine the regression predictions from the main model, \( \hat{y} \), with the direct CPS ASEC estimates, \( y \), weighting the contribution of these two components on the basis of their relative precision in order to obtain “shrinkage” estimates of the number of people in poverty by county. Where no direct estimates are available, the model-based estimates receive full shrinkage weight.

In order to produce final estimates of county poverty levels, the shrinkage estimates are scaled by state raking factors so that the sums of county poverty estimates equal the SAIPE program’s state poverty estimates in each state. Further details regarding the SAIPE program’s methodology can be found on the SAIPE program’s webpage: http://www.census.gov/hhes/www/saipe/documentation.html.

### 4.3 Experimental Models

To assess the impact of modeling Medicaid participant data, four specifications are tested.

The first specification is the SAIPE county poverty model, as presented in Section 4.1. The second is a county model with Medicaid participant data as the only predictor, and the third is a county model with both the SAIPE variables and Medicaid participant data as predictors. Two variants of this third specification are run: one including all Medicaid participants and one including only participants who have full eligibility. The fourth model includes the SAIPE regressors, except for the Food Stamp Program variable, with the Medicaid data.

The Medicaid data used in Models 2, 3a and 4 include all program participants, while the Medicaid data used in Model 3b exclude those without full eligibility. The models are summarized below:

1) The SAIPE program’s county poverty model (See Section 4.1)

2) A county model with Medicaid participant data as the only predictor:
   \[
   x_{6i} = \log (\text{total number of Medicaid participants in county } i)
   \]

3) A county model with both the SAIPE variables and Medicaid participant data as predictors:
   a) \( x_{6i} = \log (\text{total number of Medicaid participants in county } i, \text{Tally A}) \)
   b) \( x_{6i} = \log (\text{number of Medicaid participants who are eligible for the full scope of Medicaid benefits in county } i, \text{Tally B}) \)

4) A county model with the SAIPE regressors (except the Food Stamps variable) with the Medicaid data:
   \[
   x_{6i} = \log (\text{total number of Medicaid participants in county } i)
   \]

Regression results are presented in Table 1 on page 6 for each model specification.

### 4.4 Discussion of Results

The leftmost column of Table 1 displays regression diagnostics for Model 1. The SAIPE program’s county poverty model fits the data well, and the individual covariates are significant in both an economic and statistical sense. Eighty-nine percent of county-to-county variation in the available CPS ASEC poverty estimates is explained by the model.

*Model 2* also fits the data well. County-to-county variation in the Medicaid participant data alone explains a great deal of county-to-county variation in the available CPS ASEC poverty estimates. Still, fitted values based on a single data source, may not be as accurate as possible. Including information from multiple data sources, as in Model 1 and Models 3 through 5, reduces the importance of any one source, likely producing better estimates for small areas.
Table 1. Regression diagnostics for county poverty models. Dependent variable: 3-year weighted average (centered on 2000) of CPS ASEC estimated number of people in poverty for the 1,156 counties with CPS ASEC sample and non-zero estimated numbers of people in poverty in any of the three years (See Section 3.3).
Regression diagnostics for Model 3a show a positive estimated regression coefficient on the Medicaid variable that is significant at the 10% level. However, model fit is unchanged, and the estimated coefficients on the base SAIPE predictors remain close to what they are in Model 1, with the exception of the coefficient on the Food Stamp Program variable which falls. Likewise, the statistical significance of the base SAIPE predictors remains roughly the same when compared with those in Model 1, except for the Food Stamp Program variable, which becomes significant at only the 10% level.

Results from Model 3b, in which only Medicaid participants with full program eligibility are included, suggest that excluding participants who are only partially eligible decreases the predictive power of the Medicaid data in explaining county-level poverty estimates. The estimated coefficient on the Medicaid variable is no longer statistically significant when partial eligibles are excluded from the tally.

In Model 4, which excludes the Food Stamp Program variable, the estimated coefficient on the Medicaid variable is significant at the 1% level. Moreover, this estimated coefficient is similar in magnitude to the estimated coefficient on the Food Stamp Program variable in the base SAIPE production model, Model 1.

Scatter plots of residuals from models with the Medicaid variable versus from models without the Medicaid variable are very similar to one another for each model tested. Figure 2 below is a scatter plot of regression residuals from Model 1 versus regression residuals from Model 3a. Since the dependent variable has its logarithm taken prior to fitting the model, the model residuals are the difference between the log of direct survey estimates and the log of model predictions.

The largest residuals in the base SAIPE model, Model 1, are the same largest residuals in models that include Medicaid data, such as Models 3a and 4. It is not the case that the models with the Medicaid data reduce many of the large residuals.

Overall, despite Medicaid participant data having a strong individual correlation with CPS ASEC poverty estimates, I find that the Medicaid data, as modeled, do not show additional benefit for estimating CPS ASEC county poverty levels once other strong correlates with poverty, like IRS tax data, Food Stamp Program data and Census 2000 poverty estimates are included. Medicaid data may be individually important for estimating county poverty levels, but they are not found to be incrementally important.

5. Conclusion

Newly-available Medicaid participant data can be modeled to lend strength to survey-based poverty estimates. There is a strong positive correlation between CPS ASEC county poverty estimates and county tallies of Medicaid participants, and a regression of CPS ASEC poverty estimates on Medicaid participant totals suggests a strong correspondence for year 2000. Once other correlates with poverty, such as Food Stamp Program participants, tax exemptions in poverty and Census 2000 poverty estimates are considered as well, there is less remaining variation in CPS ASEC poverty estimates for the Medicaid participant data to uniquely explain. Running expanded poverty models that include the Medicaid variable, I find no evidence that the Medicaid variable has additional predictive power in estimating county poverty levels.

Medicaid participant data may be among the best available data sources in terms of county-level consistency across the country, availability over time and close relation to poverty status. However, this analysis suggests that modeling new poverty predictors, such as the Medicaid variable, may not improve the measured precision of county-level poverty estimates. Further improvement in model fit may be limited by sampling error in the survey-based dependent variable. In particular, some counties with small sample sizes have extreme poverty estimates, likely far from the “truth,” that relatively precise administrative records data, naturally, will not match.

If additional predictors, such as the Medicaid variable, do not improve measured model fit, there could still be unmeasured benefit on the precision of model
fitted values. Having a more diversified set of predictors decreases the reliance of model-based estimates on any one predictor, which may bring final estimates closer to “true” poverty levels, even if not measurably closer to direct estimates of poverty levels. This diversification benefit may be particularly important in smaller areas where there is a greater chance of extreme values in the input data.

As discussed in Section 3.1, implementation of the Medicaid program varies from state to state in terms of eligibility requirements, covered practices and neighborhood outreach. Future work might explore methods for adjusting the Medicaid data so that they are more comparable across state lines. Moreover, state fixed effects or random effects could be modeled in order to somewhat control for various statewide influences not explicitly contained in the model.

A future modeling possibility relates to the treatment of measurement error. In the poverty model’s current formulation, covariates based on administrative records are assumed to be measured without error. Alternatively, such error could be approximated and used to weight the importance of each regressor in predicting poverty levels by its relative precision (i.e., lack of measurement error). For example, if Medicaid data were known to be measured with less error than Food Stamp Program data, then, accounting for measurement error, the Medicaid variable would have more of an opportunity to be found an important predictor for estimating county poverty levels. Fisher (2003) and Fisher and Gee (2004) present an errors-in-variables model for use in the SAIPE program’s county poverty estimation.

References


