Measuring the Number of People Without Health Insurance:

A Test of a Synthetic Estimates Approach for Small Areas

Using SIPP Microdata

By

Carole Popoff
Branch Chief, Modeling & Outreach Branch
Housing and Household Economic Statistics Division
U.S. Census Bureau

D. H. Judson
Mathematical Statistician and Group Leader
Administrative Records Evaluation and Linkage Group
Planning, Research and Evaluation Division
U.S. Census Bureau

Betsy Fadali
Research Associate
University of Nevada, Reno

Paper presented at:

2001 Federal Committee on Statistical Methodology Conference

Fall, 2001

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Abstract

The need for detailed small area estimates of policy-relevant population characteristics has increased dramatically in recent years. In particular, the management and assessment of new health care initiatives such as the State Child Health Insurance Program (SCHIP) has created the need for estimates of the uninsured at the county and subcounty levels. Our ultimate goal is to develop a methodology that satisfies state and local needs and can be implemented within their level of expertise, budget, and other constraints. The goal of this paper is to test the validity of the specific estimation system we developed to address these needs.

In prior work we presented a mixed-method approach to estimating the number and percent uninsured based on a synthetic estimation technique that would satisfy the goal stated above. For this system we used: 1) a state representative population survey that captures uninsured status; and 2) county level estimates of age, race, sex and Hispanic origin (ARSH) that are constructed from state-specific fertility, mortality and migration experience. Using a synthetic technique, we thus make the link from the state level estimates to derive the county level estimates by specific ARSH categories.

This approach relies on the assumption that ARSH characteristics are adequate predictors of uninsured status. This paper tests that assumption. We use a logistic regression model to predict uninsured. We limit ourselves to ARSH variables in order to use the logistic regression results in our synthetic estimation system. Thus, strictly speaking, we are *not* developing a causal model, but instead are using maximum likelihood logistic regression techniques for "curve fitting." Though we fit the model at the individual level, we are really interested in the group-level proportions, not the individual outcomes. Finally, we generate the predicted proportions uninsured and regress the actual population proportions uninsured on the predicted proportions uninsured without an intercept to test for goodness of fit of the aggregate predicted proportion uninsured. We use the Survey of Income and Program Participation (SIPP).

The use of synthetic estimation techniques in conjunction with detailed program data illustrates the potential of such a system for performing policy-relevant demographic analysis. In this evaluation, we found that, *at the aggregate level*, ARSH characteristics generally predict the *proportion* of uninsured very well. This gives us substantial confidence in using such characteristics for a county level synthetic estimates system.

Introduction: The Need for Small Area Estimates

The National Research Council's Modernizing the U.S. Census (Edmonston and Schultze, 1995:16), a panel recommended that the "...Census Bureau work with state and local governments to enhance the quantity and frequency of small-area data." States need small area estimates (see, for example, Schirm, Zaslavsky, and Czajka, 2000), because they are mandated to serve new and ongoing demands for services while budgetary constraints limit the resources and expertise for producing reliable local information. For example, recent requests from such sources as State Health Departments and Community Development Block Grant writers highlight the continuing need for detailed, up-to-date demographic estimates. Targeting needs at this level has become even more critical as block grants have replaced earlier federally supported open-ended programs. For example, the 1996 Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA), which replaced the Aid to Families With Dependent Children Act (AFDC) allocates block grants to states rather than funding based on needs.

In this paper we examine the usefulness of a synthetic estimation methodology that will not only allow states to produce local-area estimates of the population without health care coverage, but which may also be useable to assess many policy issues. We assert that states and local entities can implement this technique within constrained budgets, timeframes and expertise in statistics or estimation techniques.

State Health Department needs

Because of rapid changes in the nature of the health care system at the state and national levels, researchers and policy makers need access reliable current information about the access to health insurance coverage and health care over time. Under Healthy People 2000 and its successor project, Healthy People 2010, the federal government outlined a series of measures for states to examine when developing health policy or public health strategies. In addition, the Centers for Medicare and Medicaid Services (CMS)[1] new program called State Children's Health Insurance Program (SCHIP), established by the Balanced Budget Act of 1997, has expanded health care coverage for children (U.S. Department of Health and Human Services, 2001). SCHIP is a capped entitlement program that requires states to develop plans that include consistent estimates of eligible recipients to determine program efficacy. Medicaid and other state initiatives have attempted to improve the data; however, in most cases the need still exists.

Focus of the Study

About 16 percent of the U.S. population was uninsured in 1997, 1998 and 1999 (U.S. Census Bureau, 2001). As stated above, some data collection efforts have been initiated. The State Systems Development Initiatives sponsored by the Health Resources and Services Administration (HRSA, 1999) targeted data collection and inventories for needs assessments, but good quality, periodic, county-level data are not available in most states. In this study, we continue prior work on the development of reliable estimation methods to estimate the number and percentage of uninsured people.

A Review of Some Techniques for Small Area Estimation

Sophisticated Techniques

Private vendors have responded to meet the need for estimates of basic demographic characteristics at low levels of geography. The American Community Survey will provide some estimates at the county level. Other sources include targeted surveys (either large surveys representative at the U.S. or state level or locally sponsored surveys); various

[1] CMS was formally called the Health Care Financing Association (HCFA).

administrative records series; and the decennial census long form (which may or may not include the variables of interest). For many of the small areas, the sample size will be small or non-existent, and using these sparse data for small area estimates results in large variances that make the estimates unreliable or simply not usable. For example, the U.S. Census Bureau's Current Population Survey is often used for state level estimates because the sampling scheme is a state-based design; however, it is not designed to be representative at lower levels of geography. County-level specialty surveys that actually are able to obtain reliable or unbiased results are prohibitively expensive. Administrative records series often have very limited detail on personal and household characteristics and are typically not designed for such work. Thus there is no simple solution to obtain the relevant information at small geographies. Often, the solution is to estimate the variable(s) of interest.

Several sophisticated estimation techniques have been tested and employed by federal agencies or consulting firms. In general, these methods are highly technical in nature; they require a sophisticated knowledge of modeling and statistical inference in addition to the construction, implementation and maintenance of elaborate estimation systems. These techniques "borrow strength" by employing either a variance components or an empirical Bayes (EB) approach to derive an indirect estimator (Harville, 1988, 1990, in Datta and Ghosh, 1991). Variations that have been tested include hierarchical Bayes (HB) or nested error regression models (see, for example, Datta and Ghosh, 1991 for a review and test of these techniques; and comments by Cressie and Kaiser, 1991; and Holt, 1991). These methods have been used, for example, to allocate federal funds for the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) (Schirm, et. al., 2001) and for allocating federal Title I funds for compensatory education in secondary schools (National Research Council, 1998, in Schirm, et. al., 2001).

The variations of EB or HB techniques are appropriate to obtain estimates of a single characteristic or a small number of closely related characteristics for each geographical area. These techniques may require the analyst to formulate and test the appropriate models for each of the different characteristics of interest. Schirm, Zaslavsky and Czajka, (2000) developed a method to produce estimates of large numbers of characteristics for each area using an elaborate re-weighting scheme. They fit a Poisson regression model to obtain an estimated prevalence for each small area for every household type in the database, the household types being defined by household characteristics. Thus a matrix of weights is produced with each household having a weight for each state in a state-based survey, for example. Since each household has a weight for every state, the entire database can be used to estimate variables of interest at the state level that increases the sample size and reduces the problem of small cell sizes. While this could theoretically be applied at the county level, no such scheme exists and to develop such a system would require enormous effort.

While all these techniques have much to recommend them, we reiterate that they require the resources and the sophisticated efforts of highly trained personnel and large organizations. The thrust of this paper is to expand on a system we developed to answer the needs of states and localities that have limited resources. For example, it is highly unlikely that state and local jurisdictions would have the expertise required to develop and test extremely sophisticated techniques to address their needs.

Drawbacks of Typical Estimation Techniques Used by States

States have typically used one or more of three techniques: state-specific or locality-specific surveys, state level data from the Current Population Survey (CPS) as an estimate for substate areas, or administrative data sets maintained at the state level. Each of these has features that make it useful for certain types of estimates, but all have significant weaknesses that preclude their use on a broader scale.

A Locally-Sponsored Special Survey Designed to Estimate the Topic-Specific Characteristics

Some states and counties have implemented topic-specific surveys to collect the relevant data. For example, the Nevada Center for Business and Economic Research conducted a statewide demographic survey to assess health insurance coverage. At a cost of approximately $93,000 the initial response rate of 10.1 percent was increased to 18.2 percent with a later telephone follow-up (Pencek, Daneshvary, and Schwer, 1998:244). Because health insurance status is highly correlated with social and economic characteristics that also correlate with survey non-response, insurance estimates based on surveys with this level of non-response are subject to great potential bias (Dillman, 1978; Tanur, 1981).

CPS-Based Estimates of Topic-Specific Characteristics

The US Census Bureau's Current Population Survey (CPS), March Supplement is widely used for comparative state-to-state characteristics (U.S. Census Bureau, 2001), and, in the mid-eighties it began being used to estimate the uninsured population (Nelson and Mills, 2001). The survey asks a series of questions as to whether each household member had insurance coverage <u>at any time</u> during the prior year by naming a fairly exhaustive list of types of coverage (e.g., Medicaid, private coverage, CHAMPUS, etc.). If the respondent provides negative responses to all of these questions, he/she is considered to be uninsured for all of the prior year. However, it is inadequate to accurately assess health insurance coverage at the sub-state level because the sampling scheme is a state-based design with the sample in each state being independent of the others (U.S. Census Bureau, 2001).

Pure Administrative Record Estimates of Topic-Specific Characteristics

Administrative data estimates are even more difficult to use as the basis of estimates for a characteristic like health insurance coverage. First, administrative data cannot provide direct information on people who are not covered, but only provide information on the number of people who accessed the various programs; therefore, presumably people covered subtracted from the total population should equal the number of people without coverage. Second, for this type of calculation to be accurate, in most cases, it is necessary to combine, sort and unduplicate administrative records from a variety of sources. However, technically, these records are *not* complete counts of the insured population making it difficult to accurately calculate an unduplicated count (for a list of potential errors, biases, and assorted limitations of administrative records data, see Judson and Popoff, 1998). Finally, while Medicaid, Medicare and insurance department enrollment records could potentially help in estimating the population covered by insurance, many legal, proprietary, and political barriers may prevent the establishment of the necessary linkages for unduplication operations. For example, privacy and confidentiality legislation often prevents this type of access. Given these conditions, administrative data are often used only to verify survey coverage (for example, see Card, Hildreth, and Shore-Sheppard, 2001).

Goals for a Small Area Estimating System To Be Useful to States for Small Area Estimates

More effective small areas estimates that can be easily and cost-effectively implemented require:

1. Developing an estimating <u>system</u> for health-related characteristics that can be replicated on a periodic basis to evaluate trends rather than single “points in time”;
2. Creating a system that provides estimates useable (and reasonable) at the county (or sub-county) level of geography;
3. Developing a system that will serve intercensal estimation needs;
4. Designing the system to be portable, i.e., useable by different states;
5. Designing the system to be expandable, i.e., useable to estimate different characteristics of interest, not tied to a particular data source or content;

6. Designing a system that can be implemented within budget and the technical constraints of state and local entities; and finally,
7. Verifying that the system is capable of providing reasonable estimates of the characteristic(s) of policy interest.

Appraisal of Goals

Each of the above data collection and estimation options (national or special surveys, administrative data, or elaborate estimation schemes) fails on at least one of the seven criteria above. For example, a single state-specific survey, representing only a point in time estimate, is not sufficient for developing a true estimating *system* nor does it offer trend analysis if not repeated on a regular basis, thus it fails goal #1 (periodic implementation, trend analysis). Similarly, using the CPS as the state-level data collection method suffices for goals #1 (estimating system), #4 (portable), and #5 (useful for many characteristics because of its broad coverage of many issues), but is not sufficient for goal #2 (county level of geography), due primarily to its cluster sampling strategy. Using a pure administrative records approach cannot address estimating people who are not in the records (such as the uninsured) but probably suffices for goal #1 (estimating system), goal #2 (county level of geography). It also fails on goals #4 and #5 (each administrative data set is specific to the program thus limiting the expandability and portability). A complex and sophisticated estimating system requires using one or several data sets as well as costly design and implementation. This type of system may be out of reach for most state and local entities thus goal #6 (ability of state and local jurisdictions to implement) is not satisfied.

Given this evaluation, we chose to develop a system based on synthetic estimation. For this system, we used two data sources; 1) the CPS, March Supplement as the state-level survey from which to derive estimates of uninsured status by ARSH characteristics;[2] and, 2) state-specific ARSH estimates derived from administrative records and produced in the usual manner. This system is replicable, portable, and relatively straightforward and cost-effective to implement within a reasonable level of expertise. It depends on an annually administered survey, can be expanded to other characteristics of interest because of the variety of survey questions, and can be localized by using specific ARSH estimates.

Attributes of a Synthetic Estimation Technique

A synthetic estimation system that ties age/race/sex/Hispanic origin (ARSH) estimates at the substate level to a household survey representative at the state level and conducted annually could potentially satisfy all seven objectives. For example, because the system updates on a yearly basis (based on estimated changes in the ARSH characteristics of the population and the use of an annually administered survey), it can be used for intercensal estimation (satisfying goal #3). The system has the potential to be portable (partially satisfying goal #4) because we use ARSH estimates that can be constructed by any state and a state-based survey that covers all states; i.e., the CPS. Also, the CPS contains information on many variables that are of policy interest (satisfying goal #5). Finally, it is cost-effective and relatively straightforward to implement and test satisfying goal #6.[3]

Figure 1 illustrates the advantage of making estimates by individual ARSH cells and is based on the following assumptions. First, break the population into two age groups, 0 to 40 and 40+; then into five age groups, 0-19, 20-39, 40-64, 64-79, 80+. Second, assume that 12

[2] The CPS is the basis of the official state-level health insurance estimates produced annually by the U.S. Census Bureau (Nelson and Mills, 2001; U.S. Census Bureau, 1996, 1997. 1998, 1999; see also, Czajka and Lewis, 1999 and Bennefield, 1996 for a comparison of health insurance status estimates using various surveys).

[3] Though, in this paper, insurance status is used as a test prototype, there is nothing specific to insurance status in our synthetic estimation techniques. The techniques could apply to any characteristic of interest for which an estimate is needed, *provided that the assumption that ARSH characteristics are good predictors of the policy variable of interest is correct*.

percent of the total population is uninsured, but that the proportion is *not* distributed uniformly across the age groups. This figure demonstrates why breaking the population into finer groups must necessarily generate more correct estimates. The dotted line represents the "true" proportion uninsured by single year of age. As can be seen in this example, using 12 percent across all age categories is dramatically incorrect for certain age groups that deviate from the average. A similar argument applies to the estimate using only two age groups, and, while using five (or more) age groups improves the estimate, without the full original age detail the estimate will not be completely correct. (See Appendix A for a more complete explanation of the synthetic technique used in previous work.)

--Insert figure 1 about here --

To summarize, the unique portions of this methodology include:

- Tying two independent sources of information together (ARSH estimates at the county level and survey estimates at a higher level of geography), and using the ARSH estimates in an attempt to take advantage of the porposed correlation between individuals' ARSH characteristics and other characteristics of interest; and
- Constructing detailed estimates at a level of geography much smaller than attempted before.

An Evaluation of the Synthetic Estimates Approach

Potential Difficulties Using Synthetic Estimation

Under certain conditions the underlying assumptions of a synthetic estimates system may be incorrect and thus might fail to provide reasonable estimates. The following is a list of the major assumptions that need further elaboration and testing to verify the validity of this method.

1. Representative surveys: A key component of this method is the estimate of $\hat{m}_{a,r,s,h}$[4], obtained from the population survey. To the extent that the survey fails to represent the population as a whole, this estimate may fail to represent the insurance experience of the respondents. Similarly, a survey with low response rates has the potential to generate *very small* cell sizes if respondents are differentially underrepresented in the survey itself. In addition, imprecise information will be generated from a survey with imprecise or untested questions.
2. Uniformity within ARSH cells: As illustrated in Figure 1, assuming that $\hat{m}_{a,r,s,h}$ is the same for every member of the particular a,r,s,h cell could cause the estimates to be biased. Breaking out $\hat{m}_{a,r,s,h}$ into ARSH-specific calculations reduces this bias. However, if the population is heterogeneous relative to the characteristic of interest (insurance status) *within* ARSH-specific cells, we would incur biases.
3. ARSH characteristics as useful predictors: The correlation between ARSH characteristics and certain policy-relevant characteristics may not be high enough to use to estimate the characteristics. For example, if there were no particular relationship between a person's ARSH characteristics and insurance status, using ARSH characteristics as a tool for estimating uninsured status provides no more information than multiplying an overall percentage of uninsured people times the total population number for the area. In multi-way crosstabulation terms, the table could be "collapsed" (Bishop, Feinberg, and Holland, 1975) across the irrelevant dimensions.

[4] See Appendix A for definition.

Given that we propose to use the CPS March Supplement in our estimation system proposition 1 is not of great concern in this paper. Propositions 2 and 3 are also of concern, and are analysed in this paper.

General Evaluation Approach

Synthetic methods are non-parametric methods using multiplication on a cell-by-cell basis. Our general purpose is to parameterize the predictor variables (ARSH characteristics). In this case, the synthetic method depends heavily on the ARSH characteristics to customize the estimates of the characteristic of interest to each local area. For example, suppose there are differential proportions of the population without health insurance among different the ARSH characteristics. If so, using ARSH to "customize" uninsured estimates for localities that have varying population ARSH mixes has the potential to produce customized estimates of the uninsured that reflect each area's unique population characteristics.

First, we fit an individual level logistic regression model to the data, using various combinations of ARSH characteristics as predictor variables. We also test the ranges upon which it is safe to collapse given that disaggregating survey responses into ARSH characteristics generates some unacceptably small cell sizes. Finally, we test the final estimated model for goodness of fit both at the individual level and the aggregate level. Results of this exercise formalize the notion that ARSH characteristics can be used as predictors of the odds of being uninsured, and hence can be used in a synthetic estimation system.

Indicators that ARSH characteristics correlate with uninsured status

Some evidence suggests there are differences in health care coverage among subsets of the population. For example, V. Wilcox-Gok (1989), in an assessment to determine the causal factors for the use of private health care coverage by the elderly, found that the presence of having private insurance was systematically related to race, sex, age and education.

The U.S. Census Bureau's 1996, 1997, 1998 and 1999 reports all show that people18 to 24 years old are less likely than any other age group to be insured for all of the preceding year. Examining race and ethnicity, the U.S. Census Bureau also reports that Hispanics are typically three times more likely and blacks are approximately twice as likely as white non-Hispanics to be without health insurance coverage. Bennefield (1998) used the U.S. Census Bureau's Survey of Income and Program Participation (SIPP), a longitudinal survey, to determine characteristics of people with and without continuous coverage and found that women were more likely to report continuous coverage.

Looking at changes in the number of uninsured over time, Holahan and Kim (2000) found that all race and ethnic groups experienced a significant increase in the percentage uninsured from 1994 to 1998. However, there were differences among the groups. Specifically, white non-Hispanics experienced an increase of 3.4 percent while Hispanics (of any race) experienced an increase of 21.5 percent and black non-Hispanics increased by 17.4 percent. Again, these and other studies suggest that there are differential proportions of people who have health insurance among the ARSH characteristics.

Methodology

The method chosen for this study is to estimate the probability of being without insurance coverage given individual ARSH characteristics using logistic regression[5]. The response variable, uninsured, is defined as people who had no health insurance coverage for an entire year or insured if they have had insurance coverage *for at least some part* of the year. This

[5] We limit ourselves to ARSH variables in order to use the logistic regression results in our synthetic estimation system. Thus, strictly speaking, we are *not* developing a causal model, but instead are using maximum likelihood logistic regression techniques for "curve fitting." Furthermore, though we fit the model at the individual level, we are really interested in the group-level proportions, not the individual outcomes.

definition allows us to replicate the way in which the state of being uninsured is defined by the U.S. Census Bureau's uninsured estimates from the CPS (Mills, 1999; Bennefield, 1996, 1997, 1998). If we were to fit an "intercept only" model, the mean response would simply be the overall odds of being uninsured. By fitting group- and age-specific dummy variables, the resulting odds ratios represent the odds of a particular population subgroup being uninsured relative to the reference group. (For example, an odds ratio of three for males of Hispanic origin would mean that people in this group would be three times as likely to be uninsured than people in the reference group.) Finally, we generate the predicted proportions uninsured and regress the actual population proportions uninsured on the predicted proportions uninsured without an intercept to test for goodness of fit of the aggregate predicted proportion uninsured.

Data

While we have designed our system to use the CPS which is representative at the state level, we have chosen to use the Survey of Income and Program Participation for this test. The first four waves of the 1996 panel are combined to represent an entire year's observations for each respondent who remained in sample for the entire year; this simulates the reference period for the CPS, March Supplement.[6] The 1996 panel had an increase in the sample size from prior panels. There are 80,923 (unweighted) cases or individuals included in the study.

Choosing the SIPP for this evaluation rather than CPS requires some explanation. It should be clarified that the CPS is the official source of estimates of the uninsured at the state level and that the SIPP is a nationally based survey, thus not appropriate for the system we have designed. However it is appropriate for this study. It was crucial that the survey minimize the inaccuracies in individual responses, recognizing that all surveys have accuracy

[6] Although, in theory, one year's responses are captured in three waves, we used portions of the forth wave to cover an entire calendar year.

deficiencies (see, for example, evaluations by Nelson and Mills, 2001; ASPE, 2001; Czajka and Lewis, 2000; Pascale, 2000; Lewis, Ellwood, and Czajka, 1998; Bennefield, 1996). Other large scale surveys also capture health insurance characteristics: The Medical Expenditures Survey (MEPS) administered by the Agency for Healthcare Research and Quality (AHRQ); the National Health Interview Survey (NHIS) administered by the National Center for Health Statistics (NCHS); the National Survey of American Families (NSAF) administered by the Urban Institute; and the Community Tracking Survey (CTS) administered by the Center for Studying Health System Change (Office of the Assistant Secretary for Planning and Evaluation [ASPE], 2001). All produce different estimates. There are several factors that are responsible for differing estimates:

- Reporting current status (estimates of the proportion of uninsured people would be inflated if people who are currently uninsured, but who were insured for some duration during the year, report their current status);
- The recall or reference period (the longer the time frame a person has to remember, the more likely he/she will misreport);
- How "insurance" is defined and how the questions are framed (whether or not a comprehensive list with explanations is given or whether single service versus comprehensive coverage is counted);
- The focus or detail of the questionnaire (whether or not health insurance status is a primary focus or simply adds to other information on well-being, for example); and,
- Misunderstanding of Medicaid coverage (recognized as least likely to be accurately reported in any survey due to lack of information on program eligibility [Czajka and Lewis, 2001; Halahan and Kim, 2000; Pascale, 2001]).

Given the list above of possible data accuracy problems SIPP ranks as a good choice. The recall period is four months instead of a prior year, thus the problem of point-in-time reporting instead of reference period reporting is minimized (Bennefield, 1996). In fact, one study concludes that MEPS and SIPP are the best sources to examine changes in status during a certain time period (ASPE, 2001), thus we are able to capture people who were insured for some part of the year using SIPP. The SIPP insurance questions are given in sufficient detail to cover the various types of insurance available and the focus is on comprehensive coverage. Also, SIPP is a person-based rather than household-based survey thus may more accurately reflect the status of each person (Bennefield, 1996).

Addressing the underreporting of Medicaid, a study by Card, Heldreth and Shore-Sheppard (2001) found that the SIPP adequately captured people who are covered by Medicaid. Their study matched Medicaid administrative records to SIPP respondents in California and found that the probability that the respondent reported Medicaid coverage when an administrative record existed indicating coverage was about 85 percent, (92 percent for children). They also report problems with the administrative records suggesting that the practice of using administrative records as *the* benchmark may not always be the best practice[7].

Model Specification and Results

First, we test the notion that ARSH characteristics are indicative of the proportions of the population within these categories having no health care coverage by tabulating simple descriptive statistics from our data set (unweighted). They strongly suggest that rates of uninsured people vary by different combinations of ARSH characteristics. However, there remains some question as to whether weighted or unweighted data more accurately support the synthetic estimation system. We therefore assessed the impact of different weighting schemes and within household, within family, and within subfamily clustering. We

[7] Given the underreporting of Medicaid coverage, we suggest that the number of people who reported no coverage may overstate the number of uninsured; however, all estimates using survey data have a similar problem.

performed the same regression analyses in this study with weighted data using two different weights, the sample design weights (alone) and the weights that take into account sample design, post-stratification, and attrition. Further, to account for within household, within family, and within subfamily covariance, we performed the same regression analyses with weights and these three kinds of nonindependent clustering, using robust methods developed by Huber, (1967) and White (1980, 1982). Though the clustering inflated the robust standard errors by a nominal amount, we found no substantive differences in the results. Therefore, for simplicity we present unweighted results for the remainder.

The unweighted proportion uninsured for Whites was lowest of the four race categories at about 8 percent; the percent of uninsured Blacks was 9.8 percent; it was 10.4 percent for American Indian / Alaskan Native (AEAN) and 11.2 percent for Asian / Pacific Islander (API). The uninsured percentage for males was 9.5 percent versus only 7.3 percent for females. Further cross-tabulations and collapsing of age and racial groups suggested some models to be tested. For example, females exhibited the same pattern of lower rates of being uninsured than males for any racial category. Hispanics of any race exhibited the highest percentage of being uninsured at 21.2 percent; the percentage for white-Hispanic, black-Hispanic and AEAN-Hispanic were 22.5 percent and 17.1 percent respectively. Asian / Pacific Islander-Hispanics' uninsured rate was 13.2 percent. These and other tabulations suggested the several models tested.

We begin by graphically describing patterns of the uninsured population. These figures represent unweighted data, and cubic spline interpolations have been overlaid on the data points. They illustrate differences in proportions uninsured by age, disaggregated by the four race categories, by Hispanic and non-Hispanic ethnicity, and male and female.

--Insert figure 2 about here.--

As can be seen in figure 2 (race), the general age *pattern* of proportions of uninsured people is similar among the four races, although the *level* and *variation* around the pattern differ. For example:

- The age *pattern* across race groups is very similar: The proportion of uninsured people generally increases to peak at about age 25, then drops off relatively slowly until age 64, when a steep drop occurs—after 64, almost no one is uninsured.
- The *level* of uninsured people varies slightly among the four race groups, although whites in general are less likely to be uninsured at every age. And,
- The group-specific proportions uninsured *vary* notably for AEAN and API groups—this is a reflection of the relatively small numbers of people in each of those groups. The white and black groups exhibit less variation around the "typical" age pattern.

This pattern suggests that it is possible to collapse black, AEAN, and API groups into an overall "other " category and compare that category to whites.

-- Insert figure 3 about here.--

As can be seen in figure 3, both groups follow the typical age pattern, with people other than white having a higher level after about age 12 or 13, and more variation in the ages greater than about 35.

--Insert figure 4 about here.--

As can be seen, figure 4 shows that Hispanics and non-Hispanics exhibit very similar, and typical, age patterns of uninsured persons. However, the Hispanic level is substantially higher at every age, and has much more variation after age 35 than the comparable non-Hispanics.

--Insert figure 5 about here.--

Finally, we can see in figure 5 an interesting interaction between age and sex. The age pattern for males and females is typical, and the levels are almost exactly the same before age 18 and after age 65. *However the levels diverge notably after age 18 and before age 50*, with males showing much higher proportions at all ages within this category. For ages 50 to 64, the proportion uninsured appears to be converging with increasing age.

All of these separate graphs lead to a common supposition: The *age pattern of the proportion of uninsured people is relatively stable* across different race, Hispanic, and gender groups. Thus, a model that uses age information can explain much of the group-level variation in levels of proportions uninsured. A final question remains: How should age information best be used? Figures 6 and 7 illustrate different options.

--Insert figure 6 about here.--

In figure 6 above, we have illustrated the population proportion uninsured by age, a cubic spline interpolation on the population, a "categorical" age model (in which we merely collapse age groups and fit an age-group-specific mean), and a "simple" age model. In the "simple" age model, we fit a cubic polynomial on age between ages 0 and 64, and a single dummy variable for ages 65 and older. After fitting parameters, we obtained the predicted values seen on the graph: An increasing probability of being without insurance coverage, peaking at about age 27, curving downward to age 64, with a barely recognizable upturn prior to age 65. The simple model does not fit the level of uninsured people between ages 0-64 very well, particularly between the ages of 15 and 40. The "categorical" model fits levels reasonably well, but at the cost of severely distorting the predicted age pattern, because there is no within-group slope reflected in the model.

In order to incorporate the best features of both models, we fit an "elaborated" age model. The goal of the "elaborated" age model is to provide a close fit for each single year of age, while keeping the number of estimated parameters to a minimum. The “elaborated” age model has the following features:

- An age-specific slope increasing from ages 0 to 15;
- A polynomial curve fit between ages 16 to 30;
- An age-specific slope decreasing from age 31 to 64; and
- A single parameter governing proportion uninsured from ages 65 on (captured in the intercept in the model fitting process).

The specific form of the regression model fit is as follows:

$$\ln\frac{p(Uninsured)}{1-p(Uninsured)} = b_o + b_1 Age0\_15 + b_2 Age16\_30 + b_3 Agex16\_30 + b_4 Age216\_30 + b_5 Age31\_64 + B_7 Agex015 + b_7 Agex31\_64$$

Where:

Uninsured = person with no insurance coverage for the entire year (1) or having insurance coverage for some time during the year (0);

Age0_15 = people zero to 15 years old as of December (1) or not (0);

Age16_30 = people 16 to 30 years old as of December (1) or not (0);

Agex1630 = an interaction term, taking the value 0 if the person is outside of ages 16 to 30, and taking their age if they are inside the age group 16 to 30 (*age * age16_30*);

Age21630 = agex1630 squared, that is, taking the value 0 if the person is outside of ages 16 to 30, and taking their age *squared* if they are inside the age group 16 to 30 (age^2 * *age16_30*);

Age31_64 = people 31 to 64 as of December (1) or not (0);

Agex015 = an interaction term, taking the value 0 if the person is outside of ages 0 to 15, and taking their age if they are inside the age group 0 to 15 *(age * age0_15*); and

Agex31_64 = an interaction term, taking the value 0 if the person is outside of ages 31 to 64, and taking their age if they are inside the age group 31 to 64 (*age * age31_64*).

The following figure shows the population proportion uninsured, the cubic spline interpolation, and the predicted values of the fitted model, by age.

--Insert figure 7 about here.--

As can be seen, the fitted values of the elaborated model fit the age pattern of proportions of uninsured people much better than either the categorical model or the simple model, at a cost of estimating seven parameters.

Model 1 ("Elaborated age model")

We shall call the "elaborated age model" Model 1. Model 1 includes no race, ethnicity, or gender specific effects—it is equivalent to assuming that *only* age makes a difference in a person's probability of being uninsured. As can be seen in figure 7, at the aggregate level model 1 generates predicted proportions uninsured that are consistent with the population as a whole. However, model 1 is fitted at the *individual* level; we present individual level logistic regression results below.

Number of Observations =	80923
LR chi^2 (7) =	2403.76
Prob > chi2 =	0
Log Likelihood =	-22128.05
Pseudo R^2 =	0.0515

	Odds Ratio	Std. Err.	z	P>|z|	Confidence Interval	
Age0_15	14.1306600	2.500667	14.97	0.0000	9.98913	19.98930
Age16_30	0.0266538	0.020274	-4.77	0.0000	0.00600	0.11836
Agex1630	1.9058820	0.125307	9.81	0.0000	1.67545	2.16801
age21630	0.9865926	0.001401	-9.51	0.0000	0.98385	0.98934
Age31_64	60.0980500	11.358960	21.67	0.0000	41.49322	87.04494
Agex015	1.0373600	0.007057	5.39	0.0000	1.02362	1.05128
Agex3164	0.9827579	0.002042	-8.37	0.0000	0.97876	0.98677

Note: The number of observations in 80,923 unweighted people in 1996; *Age0_15* is an indicator variable taking the value one for people 0 to 15 years old as of December and zero otherwise; *Age16_30* is an indicator variable taking the value one for people 16 to 30 years old as of December and zero otherwise; *Agex1630* is an interaction term, taking the value zero if the person is outside of ages 16 to 30, and taking their age if they are inside the age group 16 to 30 (age * age16_30); *Age21630* is *Agex1630* squared, that is, taking the value zero the person is outside of ages 16 to 30, and taking their age *squared* if they are inside the age group 16 to 30 (age^2 * age16_30); *Age31_64* is an indicator variable taking the value one for people 31 to 64 as of December and zero otherwise; *Agex015* an interaction term, taking the value zero if the person is outside of ages 0 to 15, and taking their age if they are inside the age group 0 to 15 (age * age0_15); and *Agex31_64* is an interaction term, taking the value zero if the person is outside of ages 31 to 64, and taking their age if they are inside the age group 31 to 64 (age * age31_64); Odds ratio is the estimated odds of being uninsured relative to the omitted group (age 65+); Std. Error is the nonrobust standard error of the estimate; z is the z-score associated with a null hypothesis that the Odds ratio is one; P>|z| is the probability, under the model, of obtaining a z-score at this value or higher by chance; confidence interval is the 95% confidence interval around the estimated Odds ratio; and Pseudo R^2 is LL[null]-LL[model])/LL[null], where LL[model] is this model's log likelihood and LL[null] is the log-likelihood of the null intercept-only model.

As an individual level predictive model only results in a pseudo R^2 of .0515 and all estimated odds ratio coefficients are significantly different from one at the conventional 5 percent level (two-tailed test). The interpretation of these estimates follows from the curve they are intended to represent: the odds ratio for the dummy variable *Age0_15* implies that people aged 0 to 15 are about an expected 14 times more likely to be uninsured than the reference group (those 65 and over), holding other effects constant. The coefficient on the *Agex015* interaction term implies that a one year increase in age results in an expected 1.04 times increase in the odds of being uninsured. (Thus, for example, a 10 year old person is 14.13 * 1.04^{10}=20.92 times more likely to be uninsured than a person in the reference group, all other effects held constant.)

Similarly, the odds ratio for the dummy variable *Age16_30* implies that people age 16 to 30 have a baseline odds ratio .027 times that of the reference group. The coefficient on *Agex1630* implies that the expected odds of being uninsured increases 1.91 times for each year of age, for those aged 16 to 30. The odds ratio for the squared term, *Age21630*, implies that the expected odds of being uninsured decrease with the square of each year of age, for those aged 16 to 30. (Note that this is as we would expect with an order-2 polynomial model—an increase in the proportion of the population who are uninsured, followed by a local maximum, followed by decrease.)

The estimated odds ratio for the dummy variable *Age31_64* indicates that the expected odds of being uninsured for a person aged 31 are 60.1 times greater than the reference group. The estimate odds ratio for the interaction term *Agex3164* indicate that the odds of being uninsured decrease past age 31, at an expected decrease of .983 per year, holding other effects constant. (Thus, for example, a 61 year old person is 60.1 * $.983^{30}$=35.93 times more likely to be uninsured than a person in the reference group, all other effects held constant.)

Model 2 - Fully Specified Model

As can be seen in earlier graphs, while the age *pattern* may be similar, it is inappropriate to assume that the *levels* of uninsured people are equivalent for non-whites versus whites, Hispanics versus non-Hispanics, and males versus females given our tabular results. In addition, the graphical analysis suggests an interaction between gender and ages 16 through 30. Model 2 also includes the interaction terms of male with the appropriate age categories already defined as follows:

$$\ln\frac{p(Uninsured)}{1-p(Uninsured)} = b_0 + b_1 Age0\_15 + b_2 Age16\_30 + b_3 Age31\_64 + b_4 Hisp + b_5 Nonwhite$$
$$+ b_6 Male + b_7 Male\_age1630 + b_8 Male\_age31\_64 + b_9 Age21630 + b_{10} Agex1630$$
$$+ b_{11} Agex015 + b_{12} Agex31\_64 + b_{13} Male\_agex1630 + b_{14} Male\_age21630$$

Where:

Uninsured = person with no insurance coverage for the entire year (1) or having coverage for some time during the year (0);

Age0_15 = people 0 to 15 years old as of December (1) or not (0);

Age16_30 = people 16 to 30 years old as of December (1) or not (0);

Age31_64 = people 31 to 64 as of December (1) nor not (0);

Hisp = people of Hispanic origin of all races (1) or not (0);

Other = Blacks, American Indian / Alaskan Natives, and Asian / Pacific Islander (1) or not (0);

Male = males of any age, race and Hispanic origin (1) or not (0);

Male_age1630 = males of any age, race and Hispanic origin inside the age group of 16 to 30 (1) or not (0);

Male_age31_64 = males of any age, race and Hispanic origin (1) or not (0) within the age group of 31 to 64;

Age21630 = agex1630 squared, that is, taking the value 0 if the person is outside of ages 16 to 30, and taking their age *squared* if they are inside the age group 16 to 30 (*age* ^{2} * *age16_30*);

Agex1630 = an interaction term taking the value 0 if the person is outside of ages 16 to 30, and taking their age if they are inside the age group 16 to 30 (*age * age16_30*);

Agex015 = an interaction term, taking the value 0 if the person is outside of ages 0 to 15, and taking their age if they are inside the age group 0 to 15 (*age* * *age0_15*); and

Agex31_64 = an interaction term, taking the value 0 if the person is outside of ages 31 to 64, and taking their age if they are inside the age group 31 to 64 (*age* * *age31_64*).

Male_agex1630= an interaction term, taking the value 0 if the person is female and/or outside ages 16 to 30 and taking their age if they are male and inside the age group 16 to 30 (*age* * *age16_30*);

Male_age21630= an interaction term, that is taking the value 0 if the person is female and/or outside of ages 16 to 30, and taking their age *squared* if they are male and inside the age group 16 to 30 (*age* * *age16_30*);

As with Model 1, we present the logistic regression results below.

Number of Observations =	80923
LR chi^2 (14) =	4164.96
Prob > chi^2 =	0.0000
Log Likelihood =	-21247.45
Pseudo R^2 =	0.0893

	Odds Ratio	Std. Err.	z	P>|z|	Confidence Interval	
Age0_15	10.887650	1.933589	13.44	0.0000	7.687148	15.420660
Age16_30	0.431951	0.491948	-0.74	0.4610	0.046345	4.025926
Age31_64	41.087450	7.888657	19.35	0.0000	28.202090	59.860060
Hisp	3.662423	0.114251	41.61	0.0000	3.445205	3.893338
Nonwhite	1.485525	0.047887	12.28	0.0000	1.394571	1.582410
Male	0.952728	0.054953	-0.84	0.4010	0.850888	1.066757
Male_age1630	0.003616	0.005523	-3.68	0.0000	0.000181	0.072175
Male_age31_64	1.324937	0.091133	4.09	0.0000	1.157835	1.516155
Age21630	0.992285	0.002149	-3.58	0.0000	0.988081	0.996506
Agex1630	1.442898	0.144432	3.66	0.0000	1.185855	1.755658
Agex3164	0.986493	0.002070	-6.48	0.0000	0.982444	0.990559
Agex015	1.044741	0.007200	6.35	0.0000	1.030724	1.058947
Male_agex1630	1.689270	0.228495	3.88	0.0000	1.295875	2.202090
Male_age21630	0.989360	0.002891	-3.66	0.0000	0.983710	0.995042

Note*:* Sample size is 80923 unweighted people in 1996; *Age0_15* is an indicator variable taking the value of one for people 0 to 15 years old as of December and zero otherwise; *Age16_30* is an indicator variable taking the value one for people 16 to 30 years old as of December and zero otherwise; *Age31_64* is an indicator variable taking the value one for people 31 to 64 as of December and zero otherwise; *Hisp* is an indicator variable taking the value one for people of Hispanic origin of all races and zero otherwise; *Other* is an indicator variable taking the value one for blacks, American Indian / Alaskan Natives, and Asian / Pacific Islanders and zero otherwise; *Male* is an indicator variable taking the value one for males of any age, race and Hispanic origin and zero otherwise; *Male_age1630* is an indicator variable taking the value one for males of any age, race and Hispanic origin inside the age group of 16 to 30 and zero otherwise; *Male_age31_64* is an indicator variable taking the value one for males of any age, race and Hispanic origin (1) within the age group of 31 to 64 or zero otherwise; *Age21630* is *Agex1630* squared, that is, taking the value zero if the person is outside of ages 16 to 30, and taking their age *squared* if they are inside the age group 16 to 30 (age ^{2} * age16_30); *Agex1630* is an interaction term taking the value zero if the person is outside of ages 16 to 30, and taking their age if they are inside the age group 16 to 30 (age * age16_30); *Agex015* is an interaction term, taking the value zero if the person is outside of ages 0 to 15, and taking their age if they are inside the age group 0 to 15 (age * age0_15); *Agex31_64* is an interaction term, taking the value zero if the person is outside of ages 31 to 64, and taking their age if they are inside the age group 31 to 64 (age * age31_64); *Male_agex1630* is an interaction term, taking the value zero if the person is female and/or outside ages 16 to 30 and taking their age if they are male and inside the age group 16 to 30 (age * age16_30); and, *Male_age21630* is an interaction term, that is taking the value zero if the person is female and/or outside of ages 16 to 30, and taking their age *squared* if they are male and inside the age group 16 to 30 (age * age16_30); Odds ratio is the estimated odds of being uninsured relative to the omitted group (age 65+); Std. Error is the nonrobust standard error of the estimate; z is the z-score associated with a null hypothesis that the Odds ratio is one; P>|z| is the probability, under the model, of obtaining a z-score at this value or higher by chance; confidence interval is the 95% confidence interval around the estimated Odds ratio; and Pseudo R^2 is LL[null]-LL[model])/LL[null], where LL[model] is this model's log likelihood and LL[null] is the log-likelihood of the null intercept-only model.

As with model 1, we can inspect this model at the individual level to examine coefficients and other regression results. In this regression, the reference group is white, non-Hispanic females aged 65 and older; the odds ratios represent deviations from the reference group. Appendix B contains a table that illustrates the contributions of parameter estimates to particular age and sex groups under this model[8].

Coefficients for the age components that are also in model 1 are very similar; thus, we focus on the new components for model 2. The estimated odds ratio of 3.66 for Hispanics (*Hisp*) indicates that Hispanics are 3.66 times more likely to be uninsured at every age group, holding all other effects constant. Similarly, the estimated odds ratio of 1.49 for other s (*Other*) indicates that other s are 1.49 times more likely to be uninsured at every age group, holding other effects constant.

The male x age dummies and terms require careful consideration. Relative to the reference group, males are a fraction less likely to be uninsured, although this coefficient is not statistically different from one at conventional levels. For the *Male_age1630* term, males begin the age 16 age group very unlikely to be uninsured, holding other effects constant. But the term on *Male_agex16_30* is 1.69, indicating that for every year of age past age 16, a male is 1.69 times more likely to be uninsured. The coefficient of .989 on the *Male_age216_30* term indicates that there is a local maximum to males' odds of being uninsured, and then the odds decrease thereafter.

Goodness of fit at the aggregate level

The results of this model can best be seen by comparing aggregate proportions uninsured with the predicted values for the eight race/Hispanic/sex groups separately. This figure is presented as figure 8.

--Insert figure 8 about here.--

[8] The authors thank Brett O'Hara of the U.S. Census Bureau, who suggested the table in Appendix B.

This figure displays the predicted proportion uninsured, in aggregate, for the eight groups separately, and the obtained proportions uninsured by age. Each group (white and other, Hispanic and non-Hispanic, males and females) has their own group-specific intercept, representing differences in level of people uninsured across the age span. For males, we have fit a male-by-age interaction term, representing the increased likelihood of males aged 16 to 30 to be uninsured, relative to females of the same age.

As can be seen in figure 8, for the first six groups (white non-Hispanic females, white non-Hispanic males, white Hispanic females, white Hispanic males, other non-Hispanic females, and other non-Hispanic males), the aggregate model fit is very reasonable, generating proportions as they are approximately found in the population. For the last two groups (other Hispanic females and other Hispanic males), small cell sizes begin to cause great variability: There are only 368 (unweighted) other Hispanic males in the data set, and only 438 (unweighted) other Hispanic females in the data set. Thus, there are many age cells with only one or a few observations.

As noted earlier in the paper, the purpose of fitting this regression model is to determine whether we can use these predicted values to support a county-level synthetic estimates system. Thus, the estimates will be used at the age/race/sex/Hispanic origin cell level, not the individual level. Since the estimated model will be used in an aggregate fashion, we submit that the key question for this method is whether the regression model generates the correct aggregate proportions uninsured, not whether it fits individual level outcomes.

In order to assess the aggregate question, we calculated the proportion uninsured in the population for each age, race, Hispanic, and sex cell, and also estimated the proportion uninsured for that cell under the modeled coefficients. We regressed the population proportion uninsured on the modeled proportion uninsured, without an intercept, and report the regression R^2, by group, as a measure of goodness of fit[9].

Group	Regression R^2
White non-Hispanic Females	.954
White non-Hispanic Males	.967
White Hispanic Females	.943
White Hispanic Males	.956
Other non-Hispanic Females	.923
Other non-Hispanic Males	.920
Other Hispanic Females	.621
Other Hispanic Males	.713

As can be seen, at the aggregate level, the model generally predicts age-race-sex-Hispanic origin specific proportions very well, with R^2 values above .90 in six cases. In the two smallest groups, the regression R^2 indicates that the model is not as predictive for those groups—there is more variability in aggregate proportions uninsured than can be explained by our modeling.

Goodness of fit at the aggregate level

Conclusions

To summarize, our goal was to evaluate a synthetic estimation technique for use by state and local entities for much needed small area estimates of policy relevant characteristics. In particular, the system would have to satisfy seven criteria; replicable, reasonable at the

[9] This regression R^2 takes the model prediction=0 as the null model when calculating R^2, which is standard for no-intercept models. In this particular case, the proportion uninsured persons is quite low for most age groups, hence prediction=0 is a reasonable null model.

substate level, satisfactory for intercensal needs, portable to other areas, expandable to a variety of characteristics, able to be implemented within local government constraints, and verifiable. Because the system we chose uses a well-designed and executed survey representative at the state level and ARSH estimates generated for each local area using administrative records and standard demographic techniques, it will be within the local entities' abilities to implement. Thus, the use of synthetic estimation techniques in conjunction with detailed program data illustrates the potential of such a system for performing policy-relevant demographic analysis.

In this evaluation, we found that, *at the aggregate level*, ARSH characteristics generally predict the *proportion* of uninsured people in the population very well. This finding gives us substantial confidence in using such characteristics for a county level synthetic estimates system.

References

Bishop, Yvonne, Feinberg, Stephen, and Holland Paul (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Bennefield, Robert L. (1998). "Who loses coverage and for how long?" Current Population Report P70-64. Washington D.C.: U.S. Department of Commerce, U.S. Census Bureau.

Bennefield, Robert L. (1997). "Health insurance coverage" Current Population Report P60-202. Washington D.C.: U.S. Department of Commerce, U.S. Census Bureau.

Bennefield, Robert L. (1996). "Health insurance coverage" Current Population Report P60-199. Washington D.C.: U.S. Department of Commerce, U.S. Census Bureau.

Bennefield, Robert L. (1996). A Comparative Analysis of Health Insurance Coverage Estimates: Data from the CPS and SIPP. Paper presented at the American Statistical Association's Joint Statistical Meeting on August 6, 1996 in Chicago, IL.

Card, David, Hildreth, Andrew, and Shore-Sheppard, Lara (2001). The measurement of Medicaid coverage in the SIPP: Evidence from California, 1990-1996. Paper to be presented at 2000-2001 HHS-ASPE/Census Bureau Development Grants Conference, September 6-7, 2001.

Czajka, John L. and Lewis, Kimball (1999). "Using national survey data to analyze children's health insurance coverage: An assessment of the issues." Mathematica Policy Research, Inc. Found on 4/17/2001 at: http://aspe.hhs.tov/health/reports/Survey%20Data.htm.

Dillman, Don (1978). Mail and Telephone Surveys: The Total Design Method. New York, Wiley.

Gonzalez, Maria Elena (1973). Use and evaluation of synthetic estimates. In American Statistical Association, Proceedings of the Social Statistics Section. Washington D.C.: American Statistical Association.

Gonzalez, Maria Elena, and Waksberg, Joseph (1973). Estimation of the error of synthetic estimates. Paper presented at the first annual meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

Gonzalez, Maria Elena, and Hoza, Christine (1978). Small-area estimation with application to unemployment and housing estimates. Journal of the American Statistical Association, 73:7-15.

Health Resources and Services Administration (1999). State Systems Development Initiative Competing Renewal Proposals for 1999. Document available from the authors.

Holahan, John, and Kim, Johnny (2000). "Why does the number of uninsured Americans continue to grow?" Health Affairs, 19(4).

Huber, P.J. (1967). The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1:221-233.

Hunt, Sandra (1997). Oregon Health Plan Medicaid Demonstration: Summary Calculation of Adjusted Per Capita Costs (Capitation Rates) and Stop-Loss Premium Requirements for Physical Health and Chemical Dependency Services for October 1997 Through September 1998. San Francisco: Coopers and Lybrand, L.L.P.

Judson, D.H., Popoff, Carole (1998). Research Use of Administrative Records. Unpublished document available from the authors.

Levy, Paul S. (1971). The use of mortality data in evaluating synthetic estimates. . In American Statistical Association, Proceedings of the Social Statistics Section. Washington D.C.: American Statistical Association.

Lewis, Kimball, Ellwood, Marilyn, and Czajka, John L. (1998). "Counting the uninsured: A review of the literature." Occasional Paper Number 8, Assessing the New Federalism. Washington D.C.: The Urban Institute.

Mills, Robert J. (1999). "Health insurance coverage" Current Population Report P60-211. Washington D.C.: U.S. Department of Commerce, U.S. Census Bureau.

National Center for Health Statistics (1977). Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. Data Evaluation and Methods Research, Vital and Health Statistics Series 2, No. 75. Hyattsville, MD: U.S. Department of Health, Education and Welfare.

National Center for Health Statistics (1979). Small Area Estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States. Data Evaluation and Methods Research, Vital and Health Statistics Series 2, No. 82. Hyattsville, MD: U.S. Department of Health, Education and Welfare.

Nelson, Charles T., Mills, Robert J. (2001). The march CPS health insurance verification question and its effect on estimates of the uninsured. U.S. Census Bureau research staff paper. Found on 8/13/2001 at: http://www.census.tov/hhes/hlthins/verif.html.

Office of the Assistant Secretary for Planning and Evaluation (2001). "Understanding Estimates of the Uninsured: Putting the Differences in Context." Washington, D.C.: U.S. Department of Health and Human Services: Found 9/19/2001 on: http://aspe.hhs.gov/health/reports/hiestimetes.htm.

Pascale, Joanne (2001). "The role of questionnaire design in Medicaid estimates: Results form an experiment." Paper resented at the Washington Statistical Society, march 21, 2001. Washington, D.C.: U.S. Department of Commerce, U.S. Census Bureau.

Reder, Stephen (1994). Synthetic Estimates of NALS Literacy Proficiencies from 1990 Census Microdata. Portland, OR: Northwest Regional Educational Laboratory.

Reder, Stephen (1997). Synthetic Estimates of Literacy Proficiencies for Small Census Areas. Available: http://www.casas.org/lit/litdata/reder.pdf, March 20, 1999.

Roemer, Mark I. (2000). Assessing the Quality of the March Current Population survey and the Survey of Income and Program Participation Income Estimates: 1990-1996. Internal Report. Washington, D.C.: U.S. Census Bureau. June 16, 2000.

Schaible, Wesley L., Brock, Dwight B., and Schnack, George A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. In American Statistical Association, Proceedings of the Social Statistics Section. Washington D.C.: American Statistical Association.

Schirm, Allen L., Zaslavsky, Alan M., and Czajka, John L. (2000). "Large Number of Estimates for Small Areas." Working Paper. Washington, D.C.: Mathematica Policy Research.

Pencek, Bruce, Daneshvary, Rennae, and Schwer, R. Keith (1998). Health Insurance Coverage of Nevadans, 1997. Las Vegas, NV: CBER.

Tanur, Judith (1981). Advances in methods for large scale surveys and experiments. In: National Science Foundation (Ed.), The Five-year Outlook on Science and Technology 1981. Washington, D.C.: Government Printing Office.

U.S. Department of Health and Human Services (2001). Children's Health Insurance Program. Washington, D.C.: Found on August 17, 2001 at http://www.hcfa.gov/init/kidssum.htm.

U.S. Census Bureau (2001). Current Population Survey Annual Demographic Survey, March Supplement. Survey Documentation found on: 8/20/2001 at http://www.bls.census.gov/cps/ads/1995/ssampdes.htm.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48:817-830.

White, H. (1982). Maximum likelihood estimation of mis-specified models. Econometrica, 50:1-25.

Wilcox-Gok, V. (1989). "Health insurance coverage among the elderly." U.S. Census Bureau Report No. 149. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Appendix A: A Brief Description of the Synthetic Estimates System

The notion of synthetic estimation has been well described by such authors as Gonzalez (1973), Gonzalez and Waksberg (1973) and Gonzalez and Hoza (1978). In the context of health-related characteristics, the National Center for Health Statistics has performed a variety of studies, both theoretical and empirical, on the usefulness of synthetic estimation techniques (Levy, 1971; National Center for Health Statistics, 1977; 1979; Schaible, Brock and Schnack, 1977). More recently, Reder (1997) developed methods for estimating literacy proficiencies for populations using the National Adult Literacy Survey, State Adult Literacy Surveys, Public Use Microdata Sample data, and 1990 Decennial Census tabulations for counties[10].

The basic synthetic estimation equation used here is:

$$\hat{x}_{a,r,s,h} = P_{a,r,s,h} \times \hat{\eta}_{a,r,s,h} .$$

Where,

$$a \in \{0,...,85+\}$$
$$r \in \{W, B, API, AI\}$$
$$s \in \{M, F\}$$
$$h \in \{H, \sim H\}$$

and

$P_{a,r,s,h}$ = the number of people of age a, race r, sex s, and ethnicity h;

[10] The approach presented here builds directly on Reder's work. Though the methods differ in particulars, the notion of using survey data linked with local area ARSH characteristics is taken directly from Reder (1994; 1997).

$\hat{m}_{a,r,s,h}$ = the proportion of people of age a, race r, sex s, and ethnicity h that have the characteristic of interest (in this case, who are uninsured), $\hat{m}_{a,r,s,h} \in [0,1]$;

$\hat{x}_{a,r,s,h}$ = the number of people of age a, race r, sex s, and ethnicity h that have the characteristic of interest (in this case, who are uninsured).

The model applies an estimated group-specific proportion who are uninsured to known age/race/sex/Hispanic numbers, to estimate the number of uninsured people of the specific group.

The simplest form of a synthetic estimate is, of course, to simply multiply an overall population by an overall proportion to get an overall number. "Dot" notation describes the method within the framework of developing insurance estimates. For any variable $y_{a,r,s,h}$, we define: $y_{a,\bullet,s,h} = \sum_r y_{a,r,s,h}$.

Other sums are similary defined if a, s, or h are "dotted," when we "dot" a subscript, it merely indicates summation over all elements of that subscript, or, using more common language, "collapsing" that margin.

Using $P_{a,r,s,h}$ as an example, the total population of an area is equivalent to "collapsing" all margins, or:

$$\text{Total Population} = P_{\bullet,\bullet,\bullet,\bullet} = \sum_a \sum_r \sum_s \sum_h P_{a,r,s,h}.$$

Multiplying the total population by some overall proportion of people uninsured in the population, is expressed in this notation as:

$$\text{Total Population Uninsured } = \hat{x}_{\bullet,\bullet,\bullet,\bullet} = P_{\bullet,\bullet,\bullet,\bullet} \times \hat{\pi}_{\bullet,\bullet,\bullet,\bullet}.$$

However, the ARSH-synthetic method proposed here does make the simple "overall" calculation above. Instead, we propose making the multiplication cell-by-cell, and then summing, rather than summing and then multiplying as above. Therefore, our estimating equation is:

$$\text{Total Population Uninsured } = \sum_{a}\sum_{r}\sum_{s}\sum_{h}\hat{x}_{a,r,s,h} = \sum_{a}\sum_{r}\sum_{s}\sum_{h}P_{a,r,s,h} \times \hat{\pi}_{a,r,s,h}.$$

Appendix B: A Guide to Interpreting the Parameter Estimates of Model 2

Model 2 contains a complex set of interaction terms intended to represent various effects. The table below, expressed in terms of the additive parameters of a non-exponentiated logit model, illustrates the impact of each term on the specified age/sex group. Because race and Hispanic groups are represented as individual dummy variables and do not interact with age and sex, we break out only age and sex in this table.

Age	**Sex**	Parameters associated with that age/sex group (expressed on the additive logit scale)
Age 0-15:		
	Male	$\beta_0 + \beta_1 + \beta_6 + \beta_{11}*Age$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
	Female	$\beta_0 + \beta_1 + \beta_{11}*Age$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
Age 16-30:		
	Male	$\beta_0 + \beta_2 + \beta_6 + \beta_7 + \beta_{10}*Age + \beta_9*Age^2 + \beta_{13}*Age + \beta_{14}*Age^2$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
	Female	$\beta_0 + \beta_2 + \beta_{10}*Age + \beta_9*Age^2$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
Age 31-64:		
	Male	$\beta_0 + \beta_3 + \beta_6 + \beta_8 + \beta_{12}*Age$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
	Female	$\beta_0 + \beta_3 + \beta_{12}*Age$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
Age 65+:		
	Male	$\beta_0 + \beta_6$ $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$
	Female	β_0 $+ \beta_4*Hisp$ $+ \beta_5*Nonwhite$