AN APPLICATION OF SMALL AREA ESTIMATION TECHNIQUES TO DERIVE STATE ESTIMATES OF INSURANCE COVERAGE FROM THE 1987 NMES

Jill J. Braden and Steven B. Cohen, AHCPR
Jill J. Braden, 2101 E. Jefferson, Suite 500, Rockville, MD 20852

KEY WORDS: Synthetic estimates

1. Introduction

The primary goal of many surveys is producing accurate national or regional estimates of the characteristics or parameters of interest. Thus, while the total sample will be widely distributed over the nation, the sample within smaller geographic units, such as states, may be small or even nonexistent in some instances. The resources needed to achieve adequate sample sizes within all states in order to make precise estimates for each one are rarely available, but the requests for such estimates are widespread and frequent.

The National Medical Expenditure Survey (NMES), conducted by the Center for General Health Services Intramural Research of the Agency for Health Care Policy and Research in 1987, was designed to produce estimates of survey statistics for the nation and for large domains including Census region and certain major subclasses of the population. At the same time, estimates of health care utilization and expenditures and of insurance coverage are among those in demand at the small area level and, more specifically, at the state level.

This paper will briefly describe three techniques which can be used to produce small area estimates. An overview of the NMES which focuses on aspects of the survey design which affect small area estimation will be presented. Each of the techniques will then be used with NMES data to: (1) produce state estimates of the civilian noninstitutionalized population which used Medicaid for health care expenses in 1987 and 2) produce state estimates of the uninsured civilian noninstitutionalized population the end of 1987. Finally, the performance of the strategies will be assessed in terms of accuracy.

2. Three Small Area Estimation Techniques

2.1 The NCHS Synthetic Estimator

The NCHS synthetic estimator is an approach formalized by the National Center for Health Statistics.[1] The basic assumption is that, within a demographic subgroup, the measure of the characteristic of interest for the small area is similar to that obtained for the nation or the Census region in which the small area is located.

Demographic information such as age, race, sex and income must be available both for the sample and for the small areas. D domains are formed by cross-classification of these demographic variables. To estimate the mean of characteristic Y for the small area ℓ, the estimate of $\bar{y}_{(d)}$ for each of the D domains is calculated from the survey data. The synthetic estimate for area ℓ is the sum of the weighted average of $\bar{y}_{(d)}$ across all domains, where the weight is the proportion of the population of small area ℓ that is in each domain. That is,

$$\bar{Y}_n^*(\ell) = \sum_{d=1}^{D} P(\ell d)\ \bar{y}(d)$$

where

$\bar{Y}_n^*(\ell)$ is the NCHS synthetic estimator of the mean for the criterion variable y in small area ℓ,

$P(\ell d)$ is the proportion of the ℓ-th area's population that belongs to domain d,

and $\bar{y}(d)$ is a national or regional survey estimate of the mean value of the criterion variable y for domain d.

2.2 The Sample-Regression Estimator

The sample-regression estimator is based on a regression equation using selected predictor (symptomatic) variables as independent variables and sample data for the variable of interest as the criterion or dependent variable. This approach is generally attributed to Ericksen.[2,3]

Estimates of the dependent variable $\bar{y}$ are computed at the primary sampling unit (PSU) level from survey data. Using symptomatic indicator data for these PSUs, a regression equation is developed to predict $\bar{y}$. That is

$$\bar{Y} = X\beta + E$$

where

$\bar{Y}$ is an n by 1 vector of values for the criterion variable in the n sampled PSUs, X is an n by (p+1) matrix containing the set of p symptomatic indicators for the n sampled PSUs and an indicator for an intercept term,

β is a (p+1) by 1 vector of regression coefficients, and

E is an n by 1 vector of stochastic errors.

The values of the symptomatic indicators for small areas are then substituted into the estimated regression equation to derive the estimates of the criterion variable for the small areas. Specifically, the sample regression estimator for small area ℓ is obtained as

$$\bar{y}_r^*(\ell) = X(\ell)\hat{\beta}$$

where

$\bar{y}_r^*(\ell)$ is the sample-regression estimator of the mean for the criterion variable in small area ℓ

$X(\ell)$ is the (p+1) vector of symptomatic information for local area ℓ

and β is the regression estimate obtained in fitting the model for the data from the sampled PSUs.

2.3. The Base Unit Estimator

Kalsbeek[4] and Kalsbeek and Cohen[5] propose an alternative strategy, the post-stratified or base unit estimator. Unlike the sample-regression estimator's assumption of an underlying linear model, this approach assumes no special functional form.

Small area ℓ, referred to in this strategy as the "target area," is divided into yet smaller areas, or "base units." For example, counties could be the base units within target areas such as states. The base units are then classified into groups of base units, for which estimates can be obtained from the survey. Using symptomatic information for the sample base units, S groups are formed using either a suitable clustering algorithm or by a minimum variance stratification method. An estimate of $\bar{y}$, the criterion variable of interest, is calculated for each of the S groups by taking a weighted average of the estimates for the sample base units that comprise each group. The estimate of the criterion variable for the s-th group (s = 1,2,...,S) is given by

$$\bar{y}(s) = \sum_{i \varepsilon s} W(i)\,\bar{y}(i)$$

where

W(i) estimates the proportion of the total population of sample base units in group s that is represented by base unit i and

$\bar{y}(i)$ is an estimate of the criterion variable for the i-th sample base unit.

The base unit estimate of the criterion variable for each small area ℓ is calculated as

$$\bar{y}_b^*(\ell) = \sum_{s=1}^{S} P(\ell s)\,\bar{y}(s)$$

where P(ℓs) is the proportion of the population of small area ℓ that is classified into group s.

3. Application of these Techniques to Produce State-Level Estimates from the 1987 National Medical Expenditure Survey

The 1987 National Medical Expenditure Survey (NMES II) provides measures of health status and estimates of insurance coverage and the use of services, expenditures, and sources of payment for the period from January 1 to December 31, 1987, for the civilian, noninstitutionalized population of the United States in the Household Survey and for the population resident in or admitted to nursing homes and facilities for the mentally retarded in the Institutional Population Component. The NMES is a research project of the Center for General Health Services Intramural Research in the Agency for Health Care Policy and Research.

The NMES II Household Survey sample is a national stratified multistage area probability sample of about 35,000 individuals in approximately 14,000 households.[6] The sample design specified that the household sample be spread over at least 100 separate areas to ensure sufficient geographic dispersion of the sample and allow for separate regional estimates. The first stage involved the selection of primary sampling units (PSUs), which were counties or groups of contiguous counties. The PSU sample represented a union of the national sample frames of Westat, Inc., and NORC and was comprised of 165 PSUs at 127 distinct sites. The number of counties in the PSUs ranged from one to eight counties. Nineteen PSUs contained counties in more than one state.[7]

3.1 Two Applications of the NCHS Synthetic Estimator

The domains for the NCHS synthetic estimator were established using two approaches; one national and one within Census region. At the national level, 36 domains were formed based on cross-classification by the following variables:

SEX	(Male, Female)
AGE	(Under 18, 18-64, 65 and older)
RACE	(White non-Hispanic, Other)
INCOME	(Family income up to 125% of the poverty line, Family income between 125 % and 200 % of the poverty line, Family income over 200 % of the poverty line).

These variables were selected based on their relationship to the criterion measure of interest and the availability of data at the local level.

Stepwise regression was used to determine the domain variables to be retained for the Census region approach. For the criterion variable related to Medicaid recipiency, the domains within the four Census regions were created based on cross-classification of INCOME and RACE, as defined above. The domains within the four Census regions were established based on cross-classification of INCOME and AGE for the criterion variable associated with being uninsured. The better the selection of demographic variables used in calculating the synthetic estimate, the smaller the error. The concern that national rates may not be sensitive to the factors operating within a given state that are associated with the criterion variable of interest led to the explicit inclusion of Census region as an additional demographic control measure for the second application of the NCHS synthetic estimator.

Within each of the domains, estimates of the criterion variables and standard errors were calculated from NMES data using the Taylor series linearization method, which takes into account the complex sample design of the survey. The proportion of each state's civilian, noninstitutionalized population, $P(\ell D)$, that belonged to each domain was determined from the March 1988 Current Population Survey. The CPS is a household sample survey conducted monthly by the Bureau of the Census to provide estimates of various characteristics of the population. The March CPS, also known as the Annual Demographic File, contains additional data on work experience, income, noncash benefits and migration for the previous year.[8, 9]

3.2 An Application of the Sample-Regression Estimator

Direct estimates of the criterion variables under consideration: 1) the percent of the population using Medicaid as a source of payment for health care expenses and 2) the percent of the population uninsured at the end of 1987 were derived from the NMES for each of the sampled PSUs. Associated standard errors were computed for the survey estimates using the Taylor series linearization method. PSUs which contained counties in more than one state were split to form state-contained sub-PSUs. If the resulting sub-PSU had fewer than 70 sampled persons responding for their full period of eligibility, the sub-PSU was not included in this estimation strategy. The term PSU is used in this paper for both PSU and sub-PSU, where applicable.

Symptomatic information considered relevant to predicting the criterion variables was abstracted from two sources. The Area Resource File (ARF) is a county-specific database which contains over 7,000 demographic and health-related variables.[10] Data extracted from the March, 1991 ARF for the counties represented in the NMES PSUs were used to construct symptomatic variables aggregated to the PSU level including, but not limited to:

Nursing home beds per 1000 65+ population 2
Percent of population residing in poverty 1 2
Percent of persons 65+ residing in poverty 1
Percent of the population on AFDC 2
Percent of population receiving General Assistance 1
Percent of population receiving SSI 1
1987 birthrate 2
1987 unemployment rate 1 2

(Variables followed by the number 1 were selected predictors for the Medicaid criterion variable and those followed by the number 2 were selected for the uninsured criterion variable.)

The State Medicaid and Insurance Regulation (SMIR) database contains state-specific information on the characteristics of state Medicaid programs' eligibility, service coverage and reimbursement policies and on state regulatory policies with respect to private health insurance activities.[11] Data extracted from the SMIR for the states represented in the NMES PSUs were used to construct symptomatic variables including,but not limited to:

AFDC monthly payment standard for family of two, July 1, 1987, in dollars 1
Indicator of regional variation in AFDC need and payment standards within the state 2
Indicator of state coverage AFDC-UP families 1
Indicator of Medicaid eligibility for aged, blind, and/or disabled persons being more restrictive than SSI 1 2

Stepwise regression was employed to determine the best symptomatic data to use in predicting each of the criterion variables. A weighted least-squares approach was taken with the weights defined as the reciprocals of the standard error of the estimates for the PSU. The proportion of variation in the criterion variables explained by the selected predictors (R^2) was .56 for the Medicaid criterion variable and .53 for the uninsured criterion variable.

Having identified the symptomatic variables needed for the sample-regression estimators, the Area Resource File (ARF) and the State Medicaid and Insurance Regulation database were used again to build a file containing these variables for all 3,079 counties in the nation. The single record representing Alaska on the ARF was excluded, as it mainly reflects the characteristics of Anchorage rather than the 23 Alaska boroughs and census areas. County estimates of the criterion variables obtained from the models specified in the regressions were aggregated within the states to produce state estimates of the percent using Medicaid as a source of payment and the percent uninsured at the end of 1987.

3.3 An Application of the Base Unit Estimator

Using symptomatic information for the base units (NMES sample PSUs), twelve post-strata groups were formed for estimation of Medicaid recipients and ten post-strata groups were formed for estimation of the uninsured. Symptomatic information for the Medicaid estimation included percent of the population in poverty, percent of the population on AFDC and the 1987 unemployment rate. Percent of the PSU population in poverty, percent of the population on AFDC and the 1987 birth rate were the basis of group formation for the estimation of the uninsured. The boudaries of the groups were determined using the cum $\sqrt{f}$ rule (Cochran, 1963), which involves the cumulation of the $\sqrt{f}$, where f is the population estimate for the PSUs in a specific percentage range, for example. The boundaries for the groups are chosen so that they create approximately equal intervals on the cum $\sqrt{f}$ scale.

Next, an estimate of the percent of the population receiving Medicaid or uninsured for each post-strata group was calculated by taking a weighted average of the estimates for the sample PSUs assigned to the group. The final estimate for each state by taking a weighted average of the averages for each post-strata group, where the weights represent the proportion of the population of state that resides in the counties that belong to the group.

4. Evaluation of the Performance of the Techniques

Data on Medicaid recipients are obtained annually by the Health Care Financing Administration (HCFA) from the State Medicaid agencies. These data include the number of recipients by type of service for 49 states, excluding Arizona, and the District of Columbia. An estimate of the noninstitutionalized recipients can be obtained by subtracting from the state total those receiving services in skilled nursing facilities and in intermediate care facilities. The estimates were then used to calculate the percent of the state noninstitutionalized population that were Medicaid recipients. These percents obtained from the HCFA data and those from the national NCHS synthetic estimator, the regional NCHS synthetic estimator, the sample-regression estimator and the base unit estimator for the twenty most populous states are presented in Summary Table 1.

The March 1988 Current Population Survey, used to produce $P(\ell d)$ for the NCHS domains, was also used to obtain state estimates of the uninsured which are deemed acceptably accurate for more populous states. The percent of the state population uninsured at the end of 1987 obtained from the CPS and from the two versions of the NCHS estimator, the sample-regression estimator and the base unit estimator were compared to these CPS estimates. (data not shown)

A comparison of the performance of the alternative small area estimation strategies for the 20 largest states was considered using the following measures:

1. the mean difference between the estimate obtained from the model specific small area estimator and a censal value (or an unbiased survey estimate),
2. the mean absolute difference between the estimate obtained from the model specific small area estimator and a censal value (or an unbiased survey estimate),
3. the relative absolute difference between the estimate obtained from the model specific small area estimator and a censal value (or an unbiased survey estimate), and
4. the standard deviations that measure the dispersion in synthetic estimates for each of the evaluation statistics, relative to a model specific small area estimator.

For the comparisons that were directed to 1987 state estimates of the civilian non-institutionalized population receiving Medicaid, the NCHS estimator that controlled for region generally out-performed the model based on national patterns. Although both estimators achieved the same mean difference in estimates when compared to the value obtained from the HCFA program statistics, the standard deviation of the difference measure derived from the model that controlled for regional characteristics was markedly lower than the national model. This pattern in greater reliability for the NCHS model with a regional control held firm for the other evaluation measures, which included the mean absolute difference and relative absolute difference in estimates when compared to the HCFA program statistics. Furthermore, the NCHS model with a regional control achieved an improvement in accuracy over the national model when comparing these evaluation measures of mean absolute difference and relative absolute difference in estimates, when compared to a censal value.

A comparison of the performance of the alternative estimators in contrast to the NCHS model with a regional control revealed the following pattern. The base unit estimator out-performed the regression estimator and the NCHS estimator on all the measures of accuracy under consideration. The base unit estimator exhibited the lowest mean difference, mean absolute difference and relative absolute difference in small area estimates when compared to a censal value. Furthermore, the NCHS model with a region control consistently out-performed the regression model in terms of accuracy. Although the regression model did not perform as well as the other estimators for the measures of accuracy under consideration, it was the best discriminator of the relative state rankings with respect to the percent of the population that were Medicaid recipients.

A composite synthetic estimator combines two or more small area estimators to, ideally, obtain the strength of each.[12] An examination of the performance of a composite estimator, which reflected an unweighted mean of the synthetic estimates derived by each of the estimators revealed an additional improvement in accuracy as measured by the mean absolute difference and mean relative absolute difference in model estimates from HCFA program statistics.

When attention was directed to the reliability of the estimators under consideration, a mixed pattern emerged. The standard deviation of the evaluation statistics that measured the absolute difference and relative absolute difference between model estimates and program statistics was lowest for the base unit estimator. Alternatively, the regression estimator was the most reliable estimator for the evaluation statistic that measured the difference between the state synthetic estimate and the value obtained from the program statistics. No additional improvement in the reliability of estimates was discerned when considering the composite estimator.

For the comparisons that were directed to 1987 state estimates of the civilian noninstitutionalized population uninsured at the end of 1987, the NCHS estimator that controlled for region outperformed the model based on national patterns with respect to the evaluation statistic that measured the mean absolute difference between the synthetic estimate and the CPS estimate. Although the performance of the NCHS estimator that controlled for region was less satisfactory for the other evaluation measures, it was consistently the more reliable estimator as evidenced by markedly lower standard deviations that characterized the respective difference measures.

A comparison of the performance of the alternative estimators in contrast to the NCHS model with a regional control revealed the following pattern. As before, the base unit estimator out-performed the regression estimator on all the measures of accuracy under consideration and was generally superior to the NCHS estimator. The base unit estimator exhibited the lowest mean absolute difference and relative absolute difference in small area estimates when compared to a more precise survey estimate obtained from the Current Population Survey. Furthermore, the NCHS model with a region control consistently out-performed the regression model in terms of accuracy. Although the regression model did not perform as well as the other estimators for the measures of accuracy under consideration, it was the best discriminator of the relative state rankings with respect to the percent of the population that were uninsured. In this setting, a composite estimator did not yield an additional improvement in accuracy over that achieved by the base unit method. No consistent pattern was noted with respect to the reliability of the alternative estimators.

5. Summary and Conclusions

This paper describes, applies and evaluates three techniques which can be used to produce small area estimates. Each of the techniques, as well as a composite estimator, was used with NMES data to produce state estimates of the civilian noninstitutionalized population in 1987 which: 1) used Medicaid for health care expenses and 2) was uninsured at the end of 1987. Finally, the performance of the strategies has been assessed in terms of accuracy and reliability.

Because synthetic estimates are biased estimates, a meaningful measure of their accuracy should reflect both sampling variability and bias.[13] The root mean square error is an appropriate measure and is proposed for future work. Beyond quantifying the accuracy, there is also an impetus to investigate outliers, rankings, and other patterns to understand and improve the performance of these strategies when applied to the National Medical Expenditure Survey.

In general, the magnitude of the level of error observed for the alternative synthetic estimation models for the measures of health insurance coverage under consideration raises serious concerns regarding their utility as potential surrogates to direct estimates. Alternatively, their capacity to serve as order statistics, that distinguish the relative ranking of states with respect to diverse measures of health insurance coverage, is evidenced by study findings.

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services and the Agency for Health Care Policy and Research is intended or should be inferred. The authors are indebted to Nancy Kieffer of Social and Scientific Systems for her programming support on this project. A copy of the Reference section and all Tables can be obtained by writing the authors.

Summary Table 1.
Preliminary Synthetic Estimates of the Percent of the State Population that Received Medicaid for the Twenty Largest States.

STATE	HCFA Program Statistics	NCHS Estimator National Model	NCHS Estimator Regional Model	Sample-Regression Estimator	Base Unit Estimator
California	13.0	9.3	8.3	8.7	8.6
New York	12.4	8.0	10.9	9.3	8.8
Texas	5.4	9.9	7.7	4.8	7.2
Florida	4.9	8.3	6.2	4.2	5.8
Pennsylvania	8.8	5.8	7.3	7.6	6.9
Illinois	8.4	7.8	9.9	8.6	7.7
Ohio	9.8	6.4	7.9	7.1	7.6
Michigan	11.7	6.5	8.2	8.4	9.5
New Jersey	6.7	5.6	7.6	4.3	7.5
North Carolina	5.6	8.0	5.9	4.2	6.4
Georgia	7.6	9.1	6.7	5.5	9.6
Virginia	5.1	6.6	4.7	3.3	6.8
Massachusetts	8.5	5.0	6.4	5.6	7.0
Indiana	4.6	6.3	7.9	3.1	6.1
Missouri	6.7	6.5	8.0	6.8	7.2
Tennessee	8.5	8.2	6.1	4.9	8.2
Wisconsin	7.4	5.5	6.8	5.8	5.1
Washington	8.1	5.7	5.8	6.3	7.1
Maryland	6.4	6.6	4.7	3.7	6.9
Louisiana	9.7	11.2	8.9	10.2	10.1

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey, 1987 - Household Survey.