Estimating the Proportion of Uninsured Persons at the County Level: Exploring the Use of Additional Covariates in a Synthetic Estimates System[1]

Carole Popoff, Brett O'Hara and D. H. Judson[2]

Key Words: Small Area Estimates, Synthetic Estimation, Logistic Regression, Hierarchical Modeling

1. Introduction

The number of people who do or who do not have health insurance coverage is a persistent question driven primarily by the needs of policy-makers and others that monitor Medicaid, Medicare and other medical assistance programs. About 16 percent of the population was uninsured in 1998 and 1999 (U.S. Census Bureau, 2000). There is a growing need for estimates at lower levels of geography than the state so program administrators can target efforts and determine program efficacy. However, these estimates are not currently available from federal statistical agencies (Popoff, Judson and Fadali, 2001).

In prior work, a synthetic estimates system using age, race, sex and Hispanic origin (ARSH) characteristics was tested using logistic regression. Group- and age-specific variables were determined to be viable predictors of health insurance status (Popoff, Judson and Fadali, 2001). The resulting odds ratios represented the odds of a particular population subgroup being uninsured relative to the reference group. In this study, we improve on the previous estimates system by modeling regional cluster effects. By using hierarchical or multilevel modeling we can add relevant information and can potentially minimize the unexplained differences between regions and increase the explanatory power of the augmented ARSH model.

2. Review of Methods Used to Produce Small Area Estimates

There is not a well-administered survey with a sufficient sample size to support direct estimates of the number of uninsured persons at a sub-state level. Estimates at the state level of geography are generated by the U.S. Census Bureau (Mills, 2001). Federal agencies primarily engage in Bayesian approaches such as hierarchical Bayes (HB) and nested error regression models (Datta and Ghosh, 1991) because the quality of data and the expertise are available. The U.S. Census Bureau's Small Area Income and Poverty Estimates program (SAIPE) provides intercensal estimates of important income and poverty statistics for states, counties and school districts (Fisher, 1997; Census, 2001). For these estimates, they use an Empirical Bayes (EB) estimation method centered on linear regression. However, there are serious problems for states or smaller entities in trying to use these techniques. Primarily, these methods are impractical for deriving large numbers of estimates across many geographical regions because each estimate is essentially unique for each area (Schirm, Zaslavsky, and Czajka, 2000). States have limited resources and expertise and often need estimates on a short timeframe making these techniques unfeasible.

3. The Pure Synthetic Method

Synthetic methods have been used for sub-state estimates, but these also have shortcomings (Judson, Popoff and Fadali, 2001; Sigmund, Popoff and Judson, 1999). The synthetic technique developed by Judson, Popoff and Fadali, needs two sources of information: 1) person-level characteristics, namely age, race, sex and Hispanic origin (ARSH); and 2) an estimate of insurance status by ARSH characteristics which is normally derived from a representative survey. While the derivation of ARSH estimates done in the normal manner is readily available, using survey data has shortcomings because the survey data used must be gathered from the specific region of interest, if possible. The shortcomings are: 1) unreliable estimates due to the small or non-existent number of cases in ARSH cells; and, 2) non-uniform distributions within ARSH cells leading to biased estimates. However, the main advantage of a synthetic system for entities with limited resources is

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

2 Carole Popoff and D. H. Judson of the U.S. Census Bureau, Brett O'Hara of the Social Security Administration.

that ARSH estimates can be easily replicated across many geographic areas, as ARSH characteristics are either available or easily developed.

This paper addresses the shortcomings of the survey-based synthetic method. Two questions frame this study. Does residing in a particular geographic region (in this case, a state) influence a person's probability of being uninsured? Can variables that differentiate dissimilar geographical regions improve the estimates? If we can account for regional differences, survey data from many regions can be combined to strengthen the uninsured estimates derived from the survey by adding more cases per cell. The focus of this work is to determine whether adding region-specific data would 1) decrease variation across regions; and, 2) increase the predictive power of the model. The technique that we use to model both region- and person-level information is a hierarchical, or multilevel, model.

4. The Multilevel Model – General Discussion

Correlation between lower level units (persons) and higher level units or clusters (families, counties, or states) needs to be explained in a single model. A multilevel model identifies factors that explain why clusters are different. By making adjustments for being in different clusters, the estimates for the lower level variables are more accurate. The correct specification would nest the individuals within the region (cluster) in which they live and the statistical algorithm would compute the correct variance (variance adjustment due to clustering). This is different than controlling for simple clustering because investigating higher level relationships for potential impacts on lower levels is not done.

Thus, we chose the multilevel method for two reasons. First, geographic regions in which individuals are "nested" may account for some of the variation among outcomes (Hox, 1995). In the full multilevel model we specify, individuals are nested in states because state-specific characteristics may affect people's ability to obtain health insurance. The goal is to make the cluster effect measured by the interclass correlation, after making model-based adjustments, equal zero. Second, this design allows the testing of whether there is explainable variation between regions; it might be possible to adjust ARSH characteristics to reflect the uniqueness of a particular region. This would allow a researcher to increase the number of cases in each ARSH cell because all states' data in a survey can be used to establish the overall rate of uninsured persons while the state characteristics, denoted by $\gamma_{01}Z_j$ and $\gamma_{01}Z_jX_{ij}$ in the equation below, form specific state estimates.

5. The Multilevel Model Used

The general linear model takes the following form:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{1j}X_{1j} + u_{0j} + \varepsilon_{ij}$$

In using this technique, the researcher can specify her beliefs about the nature of the parameters such as whether or not they have higher level interactions or have random components. For this study, we specify a two-level, random-intercept only model. The model used in this paper differs from the full multilevel model by assuming that the slopes, u_{ij}, do not have random components, although the slope differs because of region characteristics, thus we exclude the term, $u_{1j}X_{1j}$. This exclusion implies that relationships between level-one and level-two independent variables are completely deterministic. Note that the interaction term, $\gamma_{11}Z_jX_{ij}$, is included which examines cross-level interactions. The final modification to the general multilevel model is that, in this study, we have a binary dependent variable, (1 = uninsured; 0 = insured). In a generalized linear model (GLZ), the researcher is modeling a linear equation, but can specify the distribution of the dependent variable and the variance. The model we specify remains the same except that the dependent variable is linked to the predictors via a logit transformation: $Y_{ij}=\log(^{p}/_{1-p})$. The variance is derived from the binomial form of the response variable. A maximum-likelihood estimation procedure is used.

Three random-intercept models are estimated to build up to the multilevel model. A baseline model, referred to as the intercept-only model, is estimated to determine the strength of the intra-state clustering using a dummy variable to indicate each state. The next step considers only a state-based model that shows how well state-level indicator variables can explain the intercept. If state-level variables reduce the random component in the intercept (after controlling for clustering between states as in the intercept-only model), then it is worth pursuing a two-level model. The third model considers only person-level variables; it contains the same variables used in Judson, Popoff and Fadali (2001) with an additional control for simple clustering within a state. By comparing it to the baseline model, the need for modeling the clusters is shown. The last model is a two-level hierarchical, random-intercept-only model as described above. Results discussed below show that the person-level model is improved slightly by including state-level variables. The state-level variables account for some of the controllable differences between states. The random component of the intercept decreases from the person-level

model because we adjusted for state differences making them more "alike". This was the goal -- to show that differences between states *per se* are not a barrier to using other states' data (and thus it is possible to borrow strength in a synthetic system).

Several measures were used to assess the appropriateness of the model and the goodness of fit. The appropriate test for clustering is a measure of the interclass correlation: $\hat{\rho} = \hat{\mu}/(\hat{\mu} + \hat{\varepsilon})$ where $\hat{\mu}$ is the estimate of the random component and $\hat{\varepsilon}$ is the estimate of the equation variance. If the measure of clustering decreases, the level-two model has increased the explanatory power while reducing the amount of unexplained variance across regions. A second measure for the success of the multilevel model is a decrease in the size of the random component. Because the interclass correlation is calculated from the random component, the first two measures are obviously related. The goodness of fit measure is the reduction in the "deviance" measure. As the deviance goes down, the explanatory power of the model increases.

6. Data and variable selection

Data from the Survey of Income and Program Participation (SIPP) for calendar year 1996 were the source of the outcome variable and the level one (person-level) variables. The definition of an uninsured person is someone who has been without insurance coverage for the entire year. In Popoff, Judson and Fadali (2001) logistic regression was used to predict the odds that a particular population subgroup was uninsured relative to the reference group. For this analysis, the same variables from their study were used for the person-level (or level-one) model. With the exception of age, the variables are centered on the grand mean, which affects the interpretation. If a coefficient is negative, the person is less likely to be uninsured than the average uninsured rate.

For state-level (or level-two) variables, several state characteristics are worth considering. For example certain labor force characteristics, industrial mix and other proportions seem to coincide with higher or lower proportions of persons with no health insurance (Holahan and Kim, 2000). Studies suggest that part-time workers are less likely to be offered health insurance in their benefits package. Seasonal workers may or may not have health coverage. Alternatively, high state corporate tax rates may encourage companies to offer tax-deductible benefits such as health insurance.

The state-level (or level-two) indicator variables chosen for this study come from the Bureau of Labor Statistics, the U.S. Census Bureau, the Tax Administration Organization and the U.S. Department of Agriculture. Note that all state-level variables are centered on their mean value ("grand mean centered"). Some of the variables collected were not used after a correlation analysis was conducted. As expected, state poverty rates are highly collinear with the percent of workers in retail employment, seasonal work, and the unemployment rate. The percent of the state population that was Hispanic was somewhat collinear with poverty rates. Part-time work was highly collinear (with a Pearson correlation coefficient >.5) with retail employment and somewhat collinear with poverty rates. The collinearity between the percent aged 65+ and the percent of males between eighteen and thirty-five was high but not unreasonably (rho < .5). The final set of state variables that were used were state poverty rates, percent of young adult males, percent elderly, percent Hispanic and percent in part-time employment.

7. Results [3]

First, the unconditional means, or random intercept-only, model is estimated. This model provides an estimate of the unadjusted clustering between regions and individuals. We use this model to make comparisons with the other models to determine the reduction in the interclass correlation and the random effect from unexplained clustering. The next model only includes the level-two state predictors. The state-only model gives an estimate of the sensitivity of the intercept to the state-level variables that were chosen. The third model is the original model presented in Judson, Popoff and Fadali (2001) with simple state clustering added and grand mean centering of the variables as described. The last model estimated is the full multilevel model, which includes the person-level, and state-level variables with appropriate between-level interaction terms. All models are estimated with a generalized linear equation using a logit link.

The unconditional means or intercept-only model - This model does not explain any variance, it decomposes the variance into two independent components and can be used to estimate the intra class correlation ($\hat{\rho}$).

$$y_{ij} = b_{0j} + e_{ij}, \text{ where}$$

$$b_{0j} = \gamma_{00} + \mu_{0j}$$

[3] Due to lack of space, we present only the results important to the focus of this study; namely the reduction in across-state differences. Full results for all the models estimated may be obtained from the authors.

It assumes that the intercept, γ_{00}, is the same across the state clusters or a fixed coefficient, and μ_{0j} measures the residual error variation or random component among state clusters. For the estimated equation, the intercept, γ_{00}, is estimated at ≈ -1.36 and the random component, μ_{0j}, is estimated at ≈ .13*. The interclass correlation, $\hat{\rho}$, which is a measure for clustering, is estimated at ≈ 11.19 percent. This results shows that there is clustering and it is large enough to try a state level model. The deviance is estimated at ≈ 83,000. (The * indicates the coefficient is significantly different from zero at the .05 level.)

The State-Level Model - The state-level model is used to determine the average value of the intercept and adjustment factors between states that mitigate the strength of clustering. State predictor variables as described are used to predict the average value of the intercept, γ_{00}, and $\gamma_{01}Z_j$. The estimate for the mean value of the intercept remains nearly the same at -1.3713 versus –1.3614. The state level variables measure the differences between clusters that cause a state to have an intercept different from the mean. A smaller random component and a corresponding smaller interclass correlation indicate that the uncontrolled differences between states are smaller. The random component was reduced by 57 percent and the interclass correlation is also much lower; 4.9 percent versus 11.2 percent. The random component, is ≈ .05*, the interclass correlation is ≈ 4.92, and deviance is ≈83,000. These results show that a simple state-level model can successfully explain over half of the state clustering. (The * indicates the coefficient is significantly different from zero at the .05 level.)

The goodness of fit of the model was not improved. However, as noted in Singer (1998), the state-level (level two) model is meant to explain how clusters differ (to decrease u_{0j}), not necessarily to improve the fit of the model. The overall fit of the model remains unchanged because the state-level variables are explaining the state differences not captured by simple state clustering. The intercept, percent of young male adults, percent of elderly, percent of part-time workers and the poverty rate are all state-level variables that have coefficients that are significantly different from zero at the .05 level. Our results show that as an average state poverty rate increases ($\gamma_{01}Z_j$), the intercept for that state moves to a level above the overall average (γ_{00}); the state has a higher uninsured rate than the average state-uninsured rate. The converse is true for the proportion of young males, proportion of people over 65, and the proportion of part-time workers in the state. Our results show that a high proportion of persons over 65 reduces the state uninsured rates below the average as should be the case due to high Medicare coverage. However, we find the same effect for the proportion of part-time workers and proportion of young males, which is counter-intuitive and might be the result of high multicollinearity among the independent variables.

The Person-Level-Only Model - The next model includes only the person-level variables (the level-one model) controlling for simple clustering effects at the state level. The estimated random component of the intercept is lower than the intercept-only model (.11 versus .13) with a small decrease in the interclass correlation (10.2 percent versus 11.2 percent). Any change is unexpected; a change implies that individual characteristics explain clustering at the state level (reverse causation). However, the change was small and probably spurious. As expected, the person-level model drastically increases the explanatory power of the model; the measure used to capture improvement – deviance – was reduced by about 10 percent. The random component μ_{0j} is estimated at ≈.11*, the interclass correlation, $\hat{\rho}$, is estimated at ≈ 10.15, and deviance is ≈ 74,000. (The * indicates the coefficient is significantly different from zero at the .05 level.)

The parameter estimates must be considered in comparison with the reference group, white, non-Hispanic females 65 years of age and older. Consequently, most coefficients are expected to be positive since the reference group is expected to be insured due to Medicare coverage.

All of the coefficients for the cross-effects between being male and age are significantly different from zero at the .01 level. Hispanics and non-Whites have a higher likelihood of being uninsured than the reference group. In general, children, middle-aged adults, Hispanics and non-Whites are more likely to be uninsured than the reference group.[4]

The Multi-Level Model - The final model is a multilevel model (including both the person-level and the state-level variables) with interactions between the two levels. This is the most complicated model

[4] For a more complete discussion of the motivation for using logistic regression and the results from the prior study, see Popoff, Judson and Fadali, 2001, referenced herein.

because it incorporates different aspects of the likelihood of being uninsured both at the person and state levels. The differences between the states *per se* will be minimized (as measured by a reduction in the interclass correlation coefficient) and the explanatory power of the model will be maximized (as measured by percent reduction in the deviance). Results show that where the random component μ_{0j} is estimated at $\approx .06$, the interclass correlation $\hat{\rho}$ is estimated at ≈ 5.95, and deviance $\approx$ is 74,000 (this deviance is 10 percent lower than the state-level and intercept-only models).

The improvements in the random component and the interclass coefficient are smaller compared to the person-only model. This indicates that about half of the clustering has been explained with easily measurable region characteristics. There was not a measurable improvement in the overall fit of the model. Again, we include regional level variables in order to take into account variation across different regions to make region specific predictions the proportions uninsured. These regional differences do not improve the fit of the model, but they do explain the effect of regional differences on the estimates.

Examining the logit estimates shows that, for the person-level coefficients, the results are very similar to the person-level-only model. The same coefficients are significantly different from zero at the .05 level and of approximately the same magnitude. With the exception of males, the coefficients on the interaction terms of ARSH categories and the percent of young male adults in the state were significantly different from zero at the .05 level.

Interpreting cross-level interaction terms takes some care. For example, consider the interaction term males 18-35 and Hispanic ($\approx$16.5). It can be interpreted to mean that, as the percent of young male adults increases above the state average, the likelihood of percent of Hispanics that are uninsured, as a group, increases. For non-Whites and for each of the age categories, the likelihood of being uninsured decreases for states which have a higher than the average proportion of young male adults.

Using the state proportion of those in poverty as one of the terms interacted with ARSH variables proved interesting. For example, as the state's poverty rate increases above the average, the coefficient for Hispanic decreases within the state. To the extent that Hispanics are more likely to be poor than the reference group, this finding does not mean that Hispanics will have a lower incidence of uninsured rates. It does mean that a high poverty rate within the state has greater power in explaining the uninsured rate than ethnicity.

In this final model, we observe that the state characteristics of percent of young males and percent of the population in poverty have improved the overall estimate of being Hispanic. The results show that μ_{0j} is estimated at $\approx .06^{*}$, $\hat{\rho}$ is ≈ 5.95, and deviance is $\approx$74,000 (reduction > 10 percent). (The * indicates the coefficient is significantly different from zero at the .05 level.)

8. Conclusions

This paper tested the value of a multilevel model where both state (level-two) and individual characteristics (level-one) were modeled with respect to estimating the proportion uninsured. Adding state-level explanatory variables that have been reported to coincide with uninsured status is a method to reduce the effect of state clustering. Two questions framed this study: 1) whether residing in a particular geographic region (in this case, a state) affects a person's probability of being uninsured; and, 2) whether adding explanatory variables can mitigate the effects of geographic clustering. To the extent that regions are similar after model-based adjustments, their data can be combined to increase the number of cases in ARSH cells for a region. This fulfills the overarching aim of establishing a disciplined method that allows the researcher to combine all states' data from a representative survey to mitigate the problem of small cell sizes for ARSH characteristics for small area estimation.

Using a multilevel technique has proven successful at reducing the importance of clustering and providing factors to adjust state differences so that the "adjusted" state looks similar to the rest of the adjusted states. Without the multilevel model, the interaction of individual's age, race and ethnicity with the percentage of young adult males in the state or the percentage in poverty would have been missed. The influence of the state-level variables also decreased the random component of the intercept.

The current model should be expanded to include other easily gathered state characteristics. For example, industry or union composition could be included. If the correct explanatory variables are chosen in a fully specified model, the interclass correlation would approach zero and reliable adjustments for region-specific proportions of uninsured persons could be made.

A similar approach should be successful when county level data are available that includes ARSH characteristics and health insurance status. For analysts with access to county level data with the requisite information, this method is easily implemented. For example, important county level information could easily be gathered from the

Regional Economic Information System (REIS). REIS can be the foundation of the level-two model.

Other possible extensions seem less fruitful. This paper used a random-intercept model (u_{0j}) while allowing the other coefficients to vary without randomness (u_{1j} is omitted). This was done because the random effects of the ARSH variables were expected to be dependent on the random effect in the intercept. However, this assumption might be incorrect.

9. References

Datta, Gauri and Malay Ghosh (1991). "Bayesian prediction in linear models: Applications to small area estimation." *Annals of Statistics (19*), 4: 1748-1770.

Fisher, Robin (1997). Methods used for small area poverty and income estimation. Presented in the 1997 Annual Meeting of the American Statistical Association and can be found in the American Statistical Association, Proceedings of the Section on Government Statistics and Section on Social Statistics: 177-191.

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

Hox, J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Judson, Dean, Carole Popoff and Betsy Fadali (2001). *Uninsured Estimates, 1999, 2000*. For Great Basin Primary Care Association, Carson City, NV: Decision Analytics, Inc.

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

Popoff, Carole, Dean Judson, and Betsy Fadali (2001). Measuring the Number of People without Health Insurance: A Test of a Synthetic Estimates Approach for Small Areas Using SIPP Microdata. Presented at the Federal Committee on Statistical Methodology Conference, November 2001.

Schirm, Allen L., Zaslavsky, Alan M., and John Czajka (2000). State Estimates of Uninsured Children, January 1998. Washington, D.C.: Mathematica Policy Research. A report submitted to U.S. Department of Health and Human Services. At: http://aspe.hhs.gov/health/reports/stateuninsured%20Children%20(CPS)/index.htm.

Sigmund, Charles L., Carole Popoff and Dean Judson (1999). A system of synthetic estimates of health-related characteristics: Linking a population survey with local data. Paper presented at the 1999 Population Association of American Meetings, New York.

Singer, J. D. (1998) "Using SAS PROC MIXED to fit multilevel models, hierarchical models: Issues and methods." *Journal of Educational and Behavioral Statistics 23* (4): 323-55.

U.S. Census Bureau (2000). *Health Insurance Coverage.* Washington. D.C.: U.S. Department of Commerce, U.S. Census Bureau.

U.S. Census Bureau (2001). *State and County Income and Poverty Estimates: Introduction to the SAIPE Project.* Found on: 11/13/2001 at http://www.census.gov/hhes/www/saipe/nontechnod/intro.html.