Review of Experimental Estimates from the U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE) Program: Comments from the U.S. Census Bureau

Robin Fisher, Joanna Turner, and Brett O'Hara
July, 2005

The Small Area Health Insurance Estimates (SAHIE) program thanks the University of Minnesota's State Health Access Data Assistance Center (SHADAC) for their review of our experimental estimates for 2000. They have several insightful comments, and we attempt to address them here. Their review can be found at http://www.census.gov/hhes/www/sahie/feedback/shadac_review.pdf.

Model Specification Issues

The first comment is about the nomenclature of our variables, and we take the point; the suggested variables names have been adopted. The more substantive point is about the interpretation of the mean log Income to Poverty Ratio (IPR) as the log of the geometric mean IPR and the reasons for its use. The IPR is a highly skewed quantity in a population, where a few families at the higher end of the distribution may have a very high income relative to the rest of the population. The arithmetic mean on the linear scale may be highly affected by the presence of a few such families, thereby limiting its use as a predictor in the model. Taking the logs first leads to a more robust measure of the center of the distribution.

The SHADAC authors note that, besides the summary measures for the distribution we have chosen, there is at least one other, more obvious choice, namely proportions in a collection of categories defined by IPRs. While it would be possible to use proportions in IPR categories, problems of multicollinearity would make the regression coefficients highly variable and difficult to interpret. The most obvious source of multicollinearity is the necessity that the sum of the proportions in the IPR categories equal to one; this problem is mitigated somewhat when logs are taken, but is still present. Note multicollinearity is primarily a problem when the parameters are interpreted and may not be a disadvantage in estimating health insurance coverage itself.

A separate problem is that the introduction of variables that contain little information about the estimand in a regression can lead to reduced precision in the final estimates. After examining the use of categories similar to those the SHADAC authors suggest, we found that we get similar predictive power from our summary, at the cost of estimating fewer parameters.

The authors ask about the effects of the SCHIP expansion on the variances, both in the modeling and in the use of the generalized variance function (GVF) for comparative purposes. The sample expansion is in effect in the second and third year of our three-year average (that is, reference years 2000 and 2001), leading to two effects. First, one might expect the variance to go down, since the sample is larger. Second, the correlation

between the single-year direct estimate for 1999 and that for 2000 might be expected to be lower, since the sample overlap is lower; both of these would lead to lower variances for the three-year average direct estimate. The accommodation for this in the modeling is fairly automatic, since that part of the model describing the variance of the three-year direct CPS ASEC estimate is flexible enough to accommodate such changes. Regarding the use of the GVFs for comparisons, we are using the official version for the direct three-year average estimate of insurance coverage for the year 2000. This is in fact designed for use with the SCHIP expanded sample as it was implemented for that year. The point is well taken that care is necessary in using the GVFs, especially in light of some results mentioned by the authors in the review (and discussed briefly below) regarding reliability of the GVF for insurance variables.

We agree that tax data for specialized groups such as agriculture or the self-employed would likely be very helpful. Unfortunately, at this time such data are not available. We have attempted to develop similar data from other sources but issues associated with, for example, confidentiality, have proven problematic.

We hope to incorporate other kinds of predictors in the future. We should note that attempts to include the Economic Research Service (ERS) of the United States Department of Agriculture (USDA) urban influence codes for 2003 and other similar measures have yielded little benefit. We are deriving the number of persons employed in the public sector using the Quarterly Census of Employment and Wages. The decennial census and the American Community Survey are also sources for information on government employment and are under consideration.

Measurement Issues

We recognize that the definition of health insurance coverage implied by the Annual Social and Economic Supplement of the Current Population Survey (CPS ASEC) is one of several and that there is some controversy over what their characteristics are. While, for the time being, we have chosen the expectation of the CPS ASEC direct estimate as our definition of health insurance coverage, we have conducted research into the possibility of estimating the Survey of Income and Program Participation (SIPP) and CPS ASEC coverage rates simultaneously (see Fisher and Turner, 2004 at http://www.census.gov/hhes/www/sahie/pubs/fisherturnerasa04.pdf). This research may continue. Regarding some specific questions posed by the SHADAC authors in this section,

- The sample size k_i is the sum of the numbers of households in the three years of the three-year average. They are not unique, so, on average, the overlap between consecutive years is 50 percent. This incurs a correlation between the years, which would have an effect on the estimated variances, but it is taken into account, on average, by the model.
- "Eligibility" here is as defined by the Centers for Medicare and Medicaid Services (CMS), and it really means that the person was a participant in the Medicaid program in that they applied for benefits and was eligible for covered medical services. No assumption of use of benefits is implied.
- Only South and West indicator variables were used because indicators for other

regions added little to measures of fit. Initial work on the suggestion to use an indicator for border states has proven fruitful. We may incorporate it in future versions of the model.

- The comment regarding the use of generalized variance functions for comparisons to the direct CPS ASEC estimates is welcome. We have also found such comparisons to be problematic; we are investigating this issue further.

Reference

Fisher, R., and Turner, J. (2004), "Small Area Estimation of Health Insurance Coverage from the Current Population Survey's Annual Social and Economic Supplement and the Survey of Income and Program Participation", Presented at the American Statistical Association Meetings in Toronto, Canada in August 2004.