A Three-Part Model for Survey Estimates of Proportions*

Mark Bauder and Sam Szelepka†

Abstract

In the Small Area Health Insurance Estimates program (SAHIE), we produce model-based estimates of health insurance coverage for demographic groups within states and counties. We model survey estimates of proportions insured, conditional on the actual proportions. For domains with smaller sample sizes, an assumption that these survey estimates are normally distributed can be questionable because they are bounded between zero and one and have positive probabilities of being exactly zero and exactly one. This presents a difficulty for using any fully continuous distribution to model them. Because we model for finely defined demographic groups, many sample sizes are small. In addition, proportions with health insurance are often quite high. Thus, we have a large number of survey estimates that are one and some that are zero. To handle both the boundedness of the survey estimates and their probability masses at zero and one, we have developed a model we call a "three-part" model. In the three-part model, we model the probability that a survey estimate is zero, the probability that it is one, and its distribution conditional on it not being zero or one. The models for probabilities of zero and one depend on the actual proportion, the sample size, and parameters that are estimated. Conditional on not being zero or one, we assume a beta distribution. In this paper, we describe the three-part model, present results from using the model, and diagnostics of model fit.

Key Words: SAHIE, Health Insurance, Small Area Estimation

1 Introduction

Eligibility requirements for several national programs motivate a need for estimates of the population with and without health insurance in certain demographic by geographic groups. The Centers for Disease Control and Prevention (CDC) conducts cancer-screening programs for low-income, uninsured women in specified age groups and the Children's Health insurance Program (CHIP) is interested in the low-income, under-19 demographic. The U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE) program, sponsored in part by the CDC, produces estimates of the numbers and proportions insured/uninsured for demographic groups by state and county.

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views express on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

† U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233

The demographic domains for which SAHIE produces estimates are defined by cross-classifications of income, age, sex, and race categories. Here, income is categorized by Income-to-Poverty Ratio (IPR), the ratio of family income divided by the Federal Poverty Level x 100% . These demographic domains are modeled separately for 2 geographic categories, state and county. Because the domains are disjoint, the estimates can be aggregated to higher levels to get insured rates for higher level domains of interest. For states, the domains for which estimates are produced include the full cross-classification of:

- 4 race/ethnicity categories: Hispanic, White not Hispanic, Black not Hispanic, Other not Hispanic
- 2 sex categories: male, female
- 5 IPR categories: <138, 138-200, 200-250, 250-400, 400+

along with 2 sets of age categories:

- 0-17, 18-39, 40-49, 50-64
- 0-18, 19-39, 40-49, 50-64

For counties, we use a similar set of domains, but do not include a race/ethnicity category. The second set of age categories is required to get the under-19 population estimates for CHIP.

1.1 The SAHIE Model

The ultimate goal of SAHIE modeling is to produce estimates of health insurance coverage within groups defined by income, geography, and other demographics. At the level at which SAHIE models, survey estimates of the numbers in the income groups are not reliable enough due to unacceptably high variances. Accordingly, the SAHIE model has two phases:

1. The first phase estimates, for a given state/age/race/sex or county/age/sex, the proportions in each of the 5 IPR categories.
2. The second phase estimates the proportions insured within each state/age/race/sex/IPR or county/age/sex/IPR.

Each phase of the model is a multilevel or hierarchical model related closely to the 'Fay-Herriot' model, commonly used in small area estimation. In the Fay-Herriot model, the variable of interest, say θ, conditional on some parameters and predictors, follows a linear model, and a survey estimate, $\hat{\theta}$, conditional on θ and parameters is unbiased and assumes some distribution, usually taken to be normal (Rao (2003) and Fay and Herriot (1979)). The two phases of the SAHIE model are similar, but while we model the survey estimates of the proportions, it is logit transformations of the proportions that follow a normal linear model. A second difference is in how some auxiliary data are treated in the model. In SAHIE, some auxiliary data are not treated as fixed predictors, but are instead modeled in a way similar to the survey estimates (Fisher (2003) and Fisher and Gee (2004)).

The SAHIE model is fully Bayesian, and we estimate the unknown parameters using Markov Chain Monte Carlo techniques.

This paper will focus on the second phase of the model, in which we model the proportion of the domain population with health insurance and the survey estimate. In the past, we have assumed that the survey estimates of the proportions insured are normally distributed. That assumption can be questionable for domains with smaller sample sizes, since the survey estimates are bounded between zero and one, and have positive probabilities of being zero and of being one. At the level of SAHIE modeling, domains are often quite small, and as a result, survey sample sizes are often small. In addition, insured rates are typically high, especially for high-income groups. For these reasons, survey estimates of proportions insured are often one, and some are zero. Thus, the normality assumption can be especially questionable in this case. In this paper, we present an approach to modeling a survey estimate of a proportion that captures both boundedness and probability masses at zero and one.

2 Alternative Model

2.1 The 3-Part Model

We refer to our model as a "three-part" model because it is a mixture of three distributions in which we model the probability that a survey estimate is zero, the probability that it is one, and its distribution conditional on not being zero or one.

For the i^{th} state/age/race/sex/IPR or county/age/sex/IPR group let:

$$p_i = true\ proportion\ insured$$

$$\widehat{p_i} = direct\ estimate\ of\ p_i$$

$$p_i^{(0)} = \Pr(\widehat{p_i} = 0\ )$$

$$p_i^{(1)} = \Pr(\widehat{p_i} = 1\ )$$

$$S_i = sample\ size$$

Then we assume that, given the true proportion insured in the i^{th} domain, the survey estimate is distributed as:

$$\widehat{p_i} \begin{cases} = 0 & with\ probability\ p_i^{(0)} \\ = 1 & with\ probability\ p_i^{(1)} \\ \sim Beta(\alpha_i, \beta_i) & with\ probability\ 1 - p_i^{(0)} - p_i^{(1)} \end{cases} \quad (1)$$

where $\widehat{p_i} \perp \widehat{p_j}\ \forall\ i, j$ and

$$\text{logit}(p_i) = x_i^T \beta + \varepsilon_i^M \quad (2)$$

$$\varepsilon_i^M \overset{iid}{\sim} N(0, v^M), \tag{3}$$

where $p_i^{(0)}$ and $p_i^{(1)}$ are modeled partly as functions of p_i:

$$p_i^{(0)} = (1 - p_i)^{1+\zeta_0(S_i-1)} \tag{4}$$

$$p_i^{(1)} = p_i^{1+\zeta_1(S_i-1)}, \tag{5}$$

and α_i and β_i, the parameters for the Beta distribution, are determined by $p_i^{(0)}$ and $p_i^{(1)}$ along with the assumed mean and variance of $\hat{p}_i$:

$$E(\hat{p}_i) = p_i \tag{6}$$

$$var(\hat{p}_i) = \lambda_1 p_i(1 - p_i)\frac{1}{S_i^{\lambda_2}}. \tag{7}$$

The linear model in (2) and (3), the unbiasedness assumption in (6) and the form of the variance in (7) are the same as in previous SAHIE models (Bauder and Luery (2010)). The difference is in replacing a normal distribution with (1).

2.2 Modeling Probabilities of 0 and 1

The models for $p_i^{(0)}$ and $p_i^{(1)}$, the probabilities of the survey estimate of p_i being 0 and 1, are motivated in the following way. Suppose that p_i is a population proportion, and that the population is effectively infinite. If $\hat{p}_i$ is the proportion in a simple random sample of size S_i, then $\Pr(\hat{p}_i = 1) = p_i^{S_i}$ and $\Pr(\hat{p}_i = 0) = (1 - p_i)^{S_i}$. In practice, the observations in survey samples we use are not independent, and likely to be positively correlated. When this is the case, the effective sample size is smaller than the actual sample size and so survey estimates of zero and one are more probable.

In the model proposed in 2.1, we basically assume that the effective sample size is proportional to the observed sample size, S_i, with some proportionality constant, ζ, to be estimated. However, rather than letting $p_i^{(1)} = p_i^{\zeta S_i}$, we make the following correction to the models for $p_i^{(1)}$ and $p_i^{(0)}$ to ensure that $\Pr(\hat{p}_i = 1) = p_i$ and $\Pr(\hat{p}_i = 0) = 1 - p_i$ when the sample size is 1:

$$p_i^{(1)} = p_i^{1+\zeta_1(S_i-1)} \tag{9}$$

$$p_i^{(0)} = (1 - p_i)^{1+\zeta_0(S_i-1)} \tag{10}$$

where S_i is the observed sample size and $\zeta_0, \zeta_1 > 0$.

3 Evaluation of Model

We have implemented the model, and run it on data for both states and counties. Here, we present diagnostics to evaluate the model, especially the model for the probabilities of direct estimates of zero and one. We

- present parameter estimates and assess their reasonableness, and
- within groups, we calculate the model predictions of the number estimates that are one and zero, and compare that to the actual number of estimates that are one and zero.

3.1 Parameter estimates

Table 1 and Table 2 contain posterior means and variances for the parameters (ζ_0 and ζ_1 in (10) and (9)) involved in the functions for probabilities of direct estimates of zero and of one, for counties and states. For both counties and states, we allowed the parameters to differ between children and the adult age groups, and by income group. For states, we required that the parameters be the same for probabilities of estimates of zero and one, because of the small number of zero estimates.

Table 1. Posterior means and variances of parameters involved in estimated probabilities that the ACS estimate of the proportion insured is one and zero. County data.

		ζ_0 **parameter for Pr(est. = 0)**		ζ_1 **parameter for Pr(est. = 1)**	
age	**IPR**	**mean**	**std. dev.**	**mean**	**std.dev**
0-17	0-138	0.511	0.039	0.670	0.011
	138-200	0.519	0.032	0.682	0.012
	200-250	0.581	0.041	0.729	0.012
	250-400	0.497	0.048	0.754	0.012
	400+	0.454	0.041	0.803	0.017
18-39, 40-49, 50-64	0-138	1.007	0.021	0.986	0.014
	138-200	1.032	0.025	0.940	0.011
	200-250	0.965	0.028	0.925	0.010
	250-400	0.973	0.046	0.924	0.009
	400+	0.906	0.075	0.946	0.011

Source: 2009 ACS-based model, SAHIE program, U.S. Census Bureau.

Table 2. Posterior means and variances of parameters involved in the estimated probabilities that the ACS estimate of the proportion insured is one and zero. State data.

		ζ_0, ζ_1 parameter for Pr(est. = 0) and Pr(est. = 1)	
age	**IPR**	**mean**	**std. dev.**
0-17	0-138	1.098	0.160
	138-200	0.767	0.081
	200-250	0.784	0.085
	250-400	0.778	0.077
	400+	0.816	0.081
18-39, 40-49, 50-64	0-138	1.200	0.124
	138-200	0.919	0.065
	200-250	0.934	0.061
	250-400	0.925	0.067
	400+	0.924	0.060

Source: Preliminary 2009 ACS-based model, SAHIE program, U.S. Census Bureau.

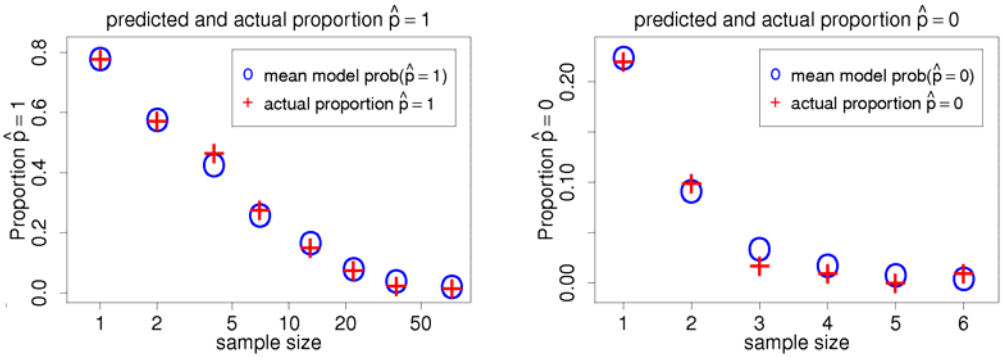
The parameter estimates are reasonable. Most are near but somewhat smaller than one. This is what we would expect. For counties, the parameters for children are smaller than for adults. This, too, is expected. Values of ζ_0 and ζ_1 can be interpreted as roughly the ratio of an "effective" sample size and what the sample size would be if the observations were independent. Smaller values reflect less independence among observations. One large source of correlation among survey estimates is the fact that whole households are in or out of sample. For something like health insurance coverage, there will be high correlation within households. But we are breaking the estimates out by age and sex. It will be more common to have two children of the same sex in a household than to have two adults of the same sex in the same age group within a household. Thus, we expect more correlation among observations within the 0-17 age group, and hence smaller ζ_0 and ζ_1 as we see for counties and for all but one income group for states.

3.2 Comparing predicted and actual estimates of zero and one

Because the probabilities that estimates are zero or one depend highly on sample size, we checked whether our model predicts the numbers of zero and one estimates well by sample size. Figure 1 and Figure 2 below contain plots of predicted and actual proportions of estimates that are one and that are zero, against sample size. Since survey estimates of one occur in areas with larger samples than do survey estimates of zero, we have put sample size into bins for $p_i^{(1)}$, and computed proportions and predicted proportions within those bins. Points within the plots in Figures 1 through three represent the following:

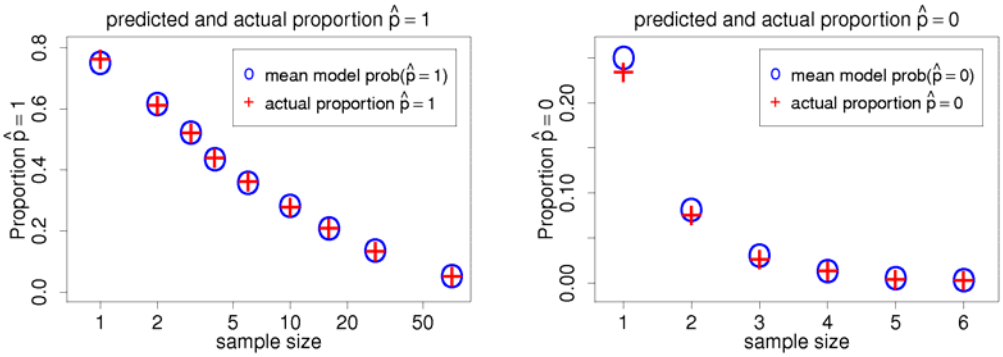
- + = actual proportion of survey estimates in a bin that are 1 (or 0)
- o = model estimate of the proportion of survey estimates in a bin that are 1 (or 0)

Figure 1. Plots of predicted and actual proportions of ACS estimates of proportion insured that are one (left) and that are zero (right), vs. sample size (in bins). States.



Source: 2009 ACS-based model, SAHIE program, U.S. Census Bureau.

Figure 2. Plots of predicted and actual proportions of ACS estimates of proportion insured that are one (left) and that are zero (right), vs. sample size (in bins). Counties.

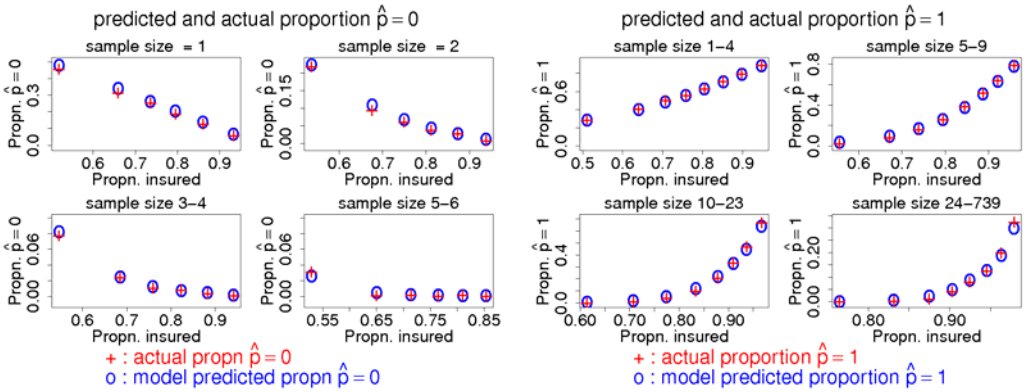


Source: 2009 ACS-based model, SAHIE program, U.S. Census Bureau.

We note that the predicted one and zero estimates closely match the actual proportions, by sample size.

The probabilities of estimates of one and zero also depend on the proportions insured, and we assessed whether our model correctly captures that dependence. Figure 3 below contains plots for counties with predicted and actual proportions of zero and one estimates plotted against the model estimates of proportion insured, divided into bins by quantile. These plots are separated by sample size group.

Figure 3. Predicted and actual proportions of ACS estimates that are zero (left) and one (right), vs. estimated proportion insured, within sample size groups. Counties.



Source: 2009 ACS-based model, SAHIE program, U.S. Census Bureau.

We see from Figure 3 that again the model predictions are very close to the actual proportions. Between the plots in Figure 1, Figure 2, and Figure 3, we see that our model for estimates of zero and one is capturing well the dependence on sample size, and the dependence on the proportions insured.

4 Summary

Previous SAHIE models assumed that the survey estimates of proportions insured are normally distributed, when in actuality they are bounded between zero and one. Because proportions insured tend to be high and many of the geographic by demographic groups are represented by very small samples, there are many groups whose survey estimates are 1, and some that are 0. The boundedness of the survey estimates together with the high densities at 0 and 1 make the current SAHIE model's assumptions questionable. The "three-part" model presented in this paper does not make the assumption of normality in the survey estimates. It instead restricts them to the interval [0,1], and models the probability that a survey estimate is 1, that it is 0, and its distribution conditional on it not being 0 or 1.

We fit the model and found that the estimates of the parameters are reasonable. Most are near one, but smaller, as expected. In counties, parameter values for children were smaller, indicating more correlation in the individual survey values. This, too, is expected.

To evaluate the performance of the "three-part" model, we compared the predicted proportion of survey estimates that are 0 and 1 with the actual proportions, plotted by bins of sample size and predicted proportion insured for states and counties. We would expect survey estimates of 0 and 1 to become less likely as sample size increases, and also depend on the predicted proportion insured. For all levels of sample size and modeled proportion insured, the predicted proportions of survey estimates that are 0 and 1 closely approximate the actual proportions. These results confirm our choice of the

functions we chose for the probabilities of estimates of one and zero, suggesting that the model is a good fit.

References

Bauder, M., Luery, D. (2010) "Small Area Estimation of Health Insurance Coverage in 2007". Available at

http://www.census.gov/did/www/sahie/methods/20062007/files/sahie_2007_technical_methodology.pdf.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", Journal of the American Statistical Association, 74, pp. 269-277.

Fisher, R. (2003), "Errors-In-Variables Model for County Level Poverty Estimation", SAIPE Working Paper,DC, U.S. Census Bureau. Available at

http://www.census.gov/did/www/saipe/publications/files/tech.report.5.pd

Fisher, R. and Gee, G. (2004), "Errors-In-Variables County Poverty and Income Models", 2004 American Statistical Association Proceedings of the Section on Government and Social Statistics. Available at

http://www.census.gov/did/www/saipe/publications/files/FisherGee2004asa.pdf.

Rao, J.N.K. (2003), Small Area Estimation, NY: Wiley.