Small Area Estimation for the Medical Expenditure Panel Survey- Insurance Component (MEPS- IC)

Steven Riesz, Bureau of the Census[1], Washington, DC 20233-6100
John Sommers, Agency for Healthcare Research and Quality, Rockville, MD 20852

Key Words: Small Area Estimation, Health Insurance, Model-assisted Estimates, Logistic Regression

Background

The Medical Expenditure Panel Survey- Insurance Component (MEPS-IC) is an annual survey of business establishments and governments sponsored by the Agency for Healthcare Research and Quality (AHRQ) and conducted by the Bureau of the Census. The survey collects data about employer-sponsored health insurance, for instance, whether or not the employer offers health insurance, what types of plans are offered, how many employees are enrolled in single and family coverage, and the amounts of the premiums that the employer and employee pay.

Private Sector Sample Design and Estimation Needs

The sample for the MEPS-IC is taken from two frames, the Census Bureau's Standard Statistical Establishment List for the private sector sample and the Census of Governments for the government sample. The work discussed in this paper deals with the private sector design and estimates only.

The private sector frame for the MEPS-IC is stratified by state and by employment size classes that are defined by a combination of company and establishment employment. Within each stratum, the frame is sorted by the first digit of an industry classification code and by establishment employment. An equal probability of selection sequential sample is drawn within each stratum. (Sommers, 1999)

The MEPS-IC was designed to produce national estimates with a relative standard error (RSE) of 1% or less, and state estimates for forty states with an RSE of 5% or less. Due to budget constraints, each year eleven states are chosen from among the twenty least populous states and the District of Columbia, and these eleven areas are given sample sizes that may lead to RSEs that are greater than 5%. State governments and other researchers and policy makers have a need, in order to make and assess state policies, for accurate estimates not only at the state level, but at sub-state levels, such as particular industries within the state, or companies of a certain size within the state. To accommodate this interest, Walkup and Sommers (2001) investigated model-assisted and composite estimates in order to improve sub-state estimates. We continue their work.

Model-Assisted Estimates

Construction of the estimates

We made model-assisted estimates of two quantities: the fraction of employees who work where health insurance is offered (fracemp), and the average single premium per enrollee (asp). (For simplicity, we will call them model estimates below.) To make these estimates, we modeled three variables for which we have survey data: whether or not an establishment offers health insurance (ins), whether or not an employee at an establishment that offers insurance enrolls in single coverage (se), and the average single premium per enrollee (aspa). Ins and se were modeled by logistic regression models and aspa was modeled by a linear regression model. All of the independent variables in the models were either discrete or were used to define discrete classes.

The estimated regression coefficients were used to predict the values of ins, se, and aspa for each establishment on the frame. The predicted values were then used to construct the estimates of interest as follows:

[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.

$$fracemp \ = \ \frac{\sum\limits_{small\ area} pins \ \cdot \ emp}{\sum\limits_{small\ area} emp}, \text{ and}$$

$$asp \ = \ \frac{\sum\limits_{small\ area} paspa \ \cdot \ pse \ \cdot pins \ \cdot \ emp}{\sum\limits_{small\ area} pse \ \cdot pins \ \cdot emp},$$

where the sums are over the establishments on the frame in a particular small area, emp is the employment of the establishment from the frame, pins is the predicted probability that the establishment offers insurance, pse is the predicted conditional probability that an employee at the establishment that offers insurance enrolls in single coverage given that insurance is offered, and paspa is the predicted average single premium per enrollee given that insurance is offered.

We modeled the three variables with several different sets of independent variables using the three models. Among the independent variables used were state, industry, employment size, county characteristics, payroll, and firm age. Weighted and unweighted models were attempted.

Variance estimation

We estimated the variances using the method of balanced half-samples (Wolter, 1985). Sixteen half-samples were defined, and the model and direct estimates were made by using only the data from the ith half-sample. The estimate of the variance of the model estimate is given by:

$$\hat{\text{var}}(\hat{m}) = \frac{\sum\limits_{i=1}^{16} (\hat{m}_i - \hat{m})^2}{16},$$

where $\hat{m}$ is the model estimate, and $\hat{m}_i$ is the model estimate obtained from the ith half-sample. A similar formula holds for variances of direct estimates.

Bias estimation

Estimates of the bias squared of the model estimates were made. These are given by:

$$bi\hat{as}^2(\hat{m}) = \max \left\{ \begin{matrix} \dfrac{\sum\limits_{i=1}^{16} (\hat{m}_i - \hat{s}_i)^2}{16} - \\ 2\dfrac{\sum\limits_{i=1}^{16} (\hat{m}_i - \hat{s}_i - (\hat{m} - \hat{s}))^2}{16}, 0 \end{matrix} \right\},$$

where $\hat{s}$ is the direct estimate, and $\hat{s}_i$ is the direct estimate using only the data from the ith half-sample. This formula is derived by first estimating the bias squared of $\hat{m}_i$, (one estimate for each half-sample), and then averaging these sixteen estimates. The first term on the right side of the equation is the average of estimates of the variance of $\hat{m}_i - \hat{s}_i$ plus the bias squared of $\hat{m}_i$, and the second term is an estimate of the variance of $\hat{m}_i - \hat{s}_i$ (the same estimate for each i.) Thus the difference of these two terms is the average of estimates of the bias squared of the $\hat{m}_i's$, and the expected value of this average is the bias squared of $\hat{m}$.

Mean squared error estimation

Estimates of the mean squared error were made. They are given by:

$$m\hat{s}e(\hat{m}) = \hat{\text{var}}(\hat{m}) + bi\hat{as}^2(\hat{m}).$$

Composite Estimates

Construction of the estimators

We also constructed composite estimates for each small area by taking a weighted average of the model based and direct estimates. These composite estimates are given by:

$$\hat{c} = \hat{w}\hat{m} + (1 - \hat{w})\hat{s},$$

where $\hat{m}$ is the model estimate, and $\hat{s}$ is the direct estimate. The optimal value of w can be derived using simple calculus as

$$w = \frac{\text{var}(s) - \text{cov}(m,s)}{\text{var}(s) + \text{var}(m) - 2\text{cov}(m,s) + bias^2(m)}, \text{ where the}$$

covariance is estimated by

$$\hat{\mathrm{cov}}(\hat{m},\hat{s}) = \frac{\sum_{i=1}^{16} (\hat{m}_i - \hat{m})(\hat{s}_i - \hat{s})}{16} .$$

In our work, if the estimate of w, which is made by replacing the quantities in the expression above by their estimates, is not between 0 and 1, then $\hat{w}$ is set equal to 0 or 1.

Mean squared error (mse) estimation

The estimate of the mse of the composite estimate has three components. The first is obtained by using the expression for the variance of the model estimate, with $\hat{m}$ and $\hat{m}_i$ replaced by $\hat{c}$ and $\hat{c}_i$, respectively. The second is the value of $\hat{w}$ squared times the bias squared of the model estimate. These are the standard portions of the mse, the variance plus the bias squared. We obtain a third term to take into account the fact that w is estimated. It has been shown in small area estimation that if parameters are estimated a term must be included for this variation also (Prassad and Rao, 1990). In our problem, we estimated the value of w for the full sample and for partial sets of half samples and then calculated the variation obtained by fixing $\hat{m}$ and $\hat{s}$ and varying the value of $\hat{w}$.

Results

One of our goals was to produce model estimates with less bias than those produced by Walkup and Sommers, 2001. We found that when we modeled ins without using weights in the model, the estimates of fracemp and asp at the state level were often biased. (We tested the estimate for bias. If a confidence interval for the difference between the model estimate and the direct estimate does not contain zero, the estimator was counted as biased.) However, when weights were used in the model, the estimates were less biased. For instance, without weights, about 70% of the state estimates of fracemp tested as having a negative bias, whereas with weights, only 8% did.

The biased estimates of fracemp produced by the unweighted model were often negatively biased. We speculate that the addition of weights to the model reduces the bias because there is a difference in the probability that an establishment offers insurance in the companies in the smallest employment class used in the model. Because of gradual changes in the chance of offering health insurance as small companies grow, our use of a categorical variable for very small companies was essentially producing an average effect for that level of effect. Since the sample has differential weights this average, which did not consider weights, did not represent the population average effect for this group of companies. With weights the effect for that level gave a better population average effect for the model. The levels of bias were reduced considerably from those obtained by Walkup and Sommers, 2001. (We should note that we tried to reduce this bias first by adding independent variables to the model, but this did not work, so we finally used the weighted models. The reason we were hesitant to use weighted models is that the weighted models are more computer intensive and tend to have more variance. Because of the use of multiple models, the sample size, and the need to predict millions of values for every half sample, we preferred to use as simple a process as possible. In the end, we spent the time on weighting because the extra bias from the unweighted models was far greater than any increase in the variance using weighted models. All results given are for weighted models.)

Comparison of values of the direct, model, and composite estimators

One of our first concerns about the small area estimates is whether they seem reasonable, since experience has shown that sometimes many of the small area estimates are just slightly different from the national mean (Ghosh and Rao, 1994). Table 1 shows the average, minimum, median, 90th percentile, and maximum values for the set of state estimates produced by the three types of estimators, for the two quantities that were estimated. One can see that for both fracemp and asp the values for any of the estimators for any of levels, average, median etc., are essentially the same. This is no surprise since the sample was designed to produce direct estimates at the state level that are usually very good. As we will see later, the model and composite estimates produced for state level estimates are of very similar quality to the direct estimates for states. No improvement is made on the direct estimates at the state level. What is reassuring about Table 1 is that the synthetic estimates have a very reasonable range. Also when checked against other available outside data, for instance, there is information about which states have the most expensive premiums, our direct and model estimates had these states as most expensive. We also calculated the correlations between the model and direct estimates within several types of small areas, and found them to be significantly correlated.

Table 2 shows the same comparisons for the estimates made for the small areas formed by crossing industries with states. Here there are several facts to notice. The synthetic estimates again have a reasonable range.

The average values of all estimates made with each of the three different estimators are very similar, however, the direct estimators have a much wider range and more extreme results. This is due, as will be seen in the next section, to the generally higher variances of the direct estimates (as opposed to the model or composite estimates) for these smaller areas, for which the sample is not able to produce stable estimates.

<u>Comparison of mean squared errors by type of estimate and small area</u>

Tables 3 and 4 show, for the two quantities estimated, the average variance estimates of the three types of estimates, the average bias squared estimates for the two synthetic estimators, and the average relative root mean squared error estimates (rrmse) for all 3 estimators for 5 sets of estimates. The average size of the groups estimated decreases as one moves from left to right on the tables, beginning with the national estimate in the leftmost column and ending with the groups created by crossing state with industry and firm size.

Both tables show similar results:

- The relative root mean squared error is generally the smallest for the composite estimator, although the difference in quality of the model and composite estimators is not very large.
- The composite, model and direct estimators have similar quality results for the national estimate and the state level estimates.
- For sub-state estimates, the model and composite estimates are better than the direct estimates and their quality relative to the direct estimates improves as the size of the areas being estimated decreases.
- The model estimate usually has the lowest variance, and the direct estimate has the highest, but the model has the highest bias, while as one would expect, the bias of the composite is less than that of the model. However, for both of the synthetic estimates, the bias is small enough so that when it is added to each estimators' variance, the average rrmse for the two synthetic estimators is still better than that of the direct estimates for the smaller areas.

Conclusions

We have created estimates of employer-sponsored insurance characteristics by predicting certain variables for each member of the Census establishment frame. This was done using multiple models to predict conditional values of insurance status given a previous status. For example, the probability of an employee having single coverage at an establishment is the probability of the establishment offering insurance times the conditional probability that an employee takes single coverage given that the establishment offers insurance. Thus, for this variable we produced models of each probability and multiplied their results to predict the overall probability that an employee at this site would have single coverage. Using these predictions and the number of employees at the site reported on the frame, we calculated the expected number of employees with single coverage at that site. Given these predictions, we use all of the predictions for establishments with a given characteristic from the frame to produce estimates for any small area desired.

Because of the complexity of the estimates, half-sample replication was used to estimate variances and biases of the results produced using these methods. Three estimators were reviewed, the direct sample estimates, the model estimates described above, and a composite estimate combining these two other estimators. Our analysis showed that the two synthetic estimators worked as well as the direct estimate for areas with sample sizes large enough to produce adequate direct estimates, such as states. However, for smaller areas the two synthetic estimators worked much better than the direct estimates, with the relative quality of the synthetic estimators versus the direct estimator increasing as the sample size of the small areas decreased.

We intend to begin providing these estimates to users in the near future.

Table 1- Some comparisons of distribution characteristics of state level estimates made using direct, model, and composite estimators

	Average	Minimum	Median	90th percentile	Maximum
Fracemp					
Direct	0.879	0.626	0.893	0.919	0.985
Model	0.881	0.700	0.887	0.920	0.987
Composite	0.880	0.651	0.888	0.918	0.986
Asp					
Direct	2,340	1,918	2,289	2,685	3,123
Model	2,329	1,871	2,292	2,682	3,109
Composite	2,334	1,899	2,292	2,682	3,109

Table 2- Some comparisons of distribution characteristics of state x industry level estimates made using direct, model, and composite estimators

	Average	Minimum	Median	90th percentile	Maximum
Fracemp					
Direct	0.838	0.085	0.889	0.983	1.000
Model	0.848	0.303	0.895	0.970	0.993
Composite	0.843	0.191	0.895	0.975	0.996
Asp					
Direct	2,340	846	2,296	2,806	5,419
Model	2,329	1,764	2,295	2,633	3,198
Composite	2,308	1,106	2,286	2,636	4,050

Table 3- Selected characteristics of three types of estimators for several groups of estimates of the probability that an establishment offers health insurance

	National	State	State x size	State x industry	State x industry x size
Variance (direct)	< 0.00001	0.00061	0.00440	0.01817	0.12536
Variance (model)	< 0.00001	0.00028	0.00109	0.00062	0.00149
Variance (composite)	< 0.00001	0.00039	0.00135	0.00158	0.00348
Bias^2 (model)	0.00001	0.00040	0.00225	0.00387	0.00591
Bias^2 (composite)	< 0.00001	0.00005	0.00017	0.00037	0.00064
Rrmse (direct)	0.186	2.295	7.687	13.660	37.518
Rrmse (model)	0.457	2.143	6.836	5.193	9.137
Rrmse (composite)	0.185	1.864	4.793	4.680	8.210

Table 4- Selected characteristics of three types of estimators for several groups of estimates of the average single premium

	National	State	State x size	State x industry	State x industry x size
Variance (direct)	179	9,556	86,271	324,203	970,875
Variance (model)	170	8,625	10,641	13,011	14,195
Variance (composite)	179	8,672	13,666	35,286	41,344
Bias^2 (model)	194	225	27,730	57,893	122,166
Bias^2 (composite)	< 1	28	2,110	6,285	15,416
Rrmse (direct)	0.576	3.658	8.377	15.398	30.356
Rrmse (model)	0.826	3.565	6.147	7.665	9.534
Rrmse (composite)	0.576	3.503	4.518	6.232	7.905

References

Ghosh M, and Rao JNK (1994). Small area estimation: an appraisal. Statistical Sciences, Vol. 9, No.1, 55-93.

Prasad, NGN and Rao, JNK (1990). The estimation of the mean squared error of small area estimators, Journal of the American Statistical Association, 85, 163-171.

Sommers JP. List sample design of the 1996 Medical Expenditure Panel Survey Insurance Component. Rockville (MD): Agency for Health Care Policy and Research; 1999. MEPS Methodology Report No. 6. AHCPR Pub. No. 99-0037.

Walkup M and Sommers JP (2001). Small group estimation for the Medical Expenditure Panel survey - Insurance Component**,** Proceedings of the Section on Survey Statistics. American Statistical Association.

Wolter K. (1985). Introduction to variance estimation. New York: Springer-Verlag.