SMALL GROUP ESTIMATION FOR THE MEDICAL EXPENDITURE PANEL SURVEY - INSURANCE COMPONENT

John Sommers, Agency for Healthcare Research and Quality[1]
Michael Walkup, Bureau of the Census[2]

Key Words: small area estimation, health insurance, composite estimates, model-based estimates, logistic regression

Background

The Medical Expenditure Panel Survey - Insurance Component (MEPS - IC) is an annual survey of business establishments(single locations) sponsored by the Agency for Healthcare Research and Quality(AHRQ) and conducted by the Bureau of the Census. Data related to employer sponsored health insurance, such as, number enrolled, plans offered and their characteristics, premiums, contributions and establishment characteristics(size, industry, etc.) are collected. Information from the MEPS-IC is used in a variety of ways. Among these are:
(1) monitoring the insurance offerings of businesses, including their charges and characteristics and (2) production of estimates of total spending on health insurance used in calculation of the Gross Domestic Product .

Because most health insurance law is made at the State level and many States are in the process of implementing reforms to widen health insurance coverage and benefits and/or to control costs, the first use listed requires that estimates be available at the State level.

Sample Design and Estimation Needs

Due to the need for State information, the MEPS-IC sample design allows for a minimum sample in 40 States each year. This is the number of States AHRQ feels that it can support with the budget for the survey, while allowing enough sample in the selected States to provide a 5% relative standard error for selected State level estimates. The 20 most populous States are provided with at least the minimum sample every year. The remaining States and the District of Columbia have sample sizes that vary from year to year on a 4 year rotational basis. Each is guaranteed adequate sample for from 1 to 3 years within each 4 year cycle. (Sommers, 1999 and Insurance Component, 2001).

The sample also contains an oversample of establishments, relative to their share of total employment, from firms with less than 50 employees. This was done because much of the focus of health insurance reform activity in both the State and Federal governments is directed toward this group.

AHRQ has had numerous requests from individual States for other substate estimates. Generally, there is a desire for estimates of premiums, enrollments, percent of persons offered and enrolled by industry and size within States. There have been less frequent inquiries for this data by type of plan, as well as, premium and contribution percentiles.

Although, many of these estimates, such as, estimates for large industries within a State, have relative errors around 10% (AHRQ, 2001), due to the importance of these estimates, there is a strong desire to improve the estimates already produced and to allow for more detailed estimates for smaller groups within States. There is also a desire to produce estimates for the 11 States which currently have inadequate sample sizes.

Due to the importance of this type of estimate, AHRQ and the Census Bureau have begun studies to determine the viability and possible quality of various small area estimation techniques to produce results for smaller cells.

This study focused on three sets of estimates, (1)percent of enrollees who selected single coverage, (2)average premium for single coverage and (3) average employee contribution for single coverage. Due to the method used for model based estimates, these three variables produce a natural cluster that can be estimated in a single process. The study also focused only on States for which AHRQ has currently committed to provide estimates. The analysis considers only 40 States. For each set of estimates two types of estimators were tested, a modeling estimator and a composite type estimator.

Model Based Estimates

The first set of estimates made were created by predicting values for each establishment on the frame.

---

[1] The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

[2] This paper reports the results of research and analysis undertaken by the census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

To do this several models were run using the sample, and expected values of several variables were produced. These were then used to make estimates of various totals and ratios of these totals for subsets of the frame. Following is a brief description and example of the process.

Suppose one wished to estimate the average premium for a specific group or area, such as, a particular industry within a State. This would be

$$\text{average premium} = \sum tp_i \Big/ \sum te_i$$

where tp_i and te_i are the predicted(expected) total premium and total enrollees at the ith establishment and each sum is taken over all establishments in the cell.

To arrive at these predictions the values can be broken into parts. One can think of tp as:

te = employment*(probability that the establishment offers health insurance)*(predicted percent of employees that would take single coverage given an establishment offers health insurance).

tp = te*(the predicted premium for that establishment given it offers health insurance)

The value of the employment is available on the frame for each establishment. Each of the other parts are modeled separately.

The probability that an establishment offers health insurance was developed using a logistic regression based on the entire data set. The other two parts were developed using linear regressions run only on the sets of establishments that offer health insurance. This multiple model approach was used because there is a natural progression to each establishment value. A single model did not suffice to predict the establishment values due to the large number of cases in the sample that did not offer health insurance and thus have zero values for the premium. The large number of zeroes led to poor direct predictions if a single model of the sample values were used.

Because of the size of the frame and the processing required, the models used were very simple for this process. All models used only categorical independent variables which included state, industry group, firm age, firm size, whether the establishment was part of a multi-unit firm, category of the average payroll and interaction of industry and size of firm. Not all variables were used for all models.

We tried to use weighted models using the sampling weights. The weighted runs took much more processing time and were dropped for the models of the probability of offering health insurance and the percent taking single coverage. The unweighted results were not significantly different than the weighted models.

There were some differences between weighted and unweighted results for the two variables, 'single premium' and 'single contribution'. We speculate that because these variables, which measure a level value, rather than a ratio, were affected differently by other area related variables not included in the current models. One candidate is county, which was not used for these trial runs. Given the initial results, county or county charcteristics will be considered in future work.

It is important to note that all predictions were made using fixed effects models rather than considering geographic variables as random effects as has been done in other work. (Battese and Fuller, 1987 and Ghosh and Rao, 1994). This approach was used because the sample design of the MEPS-IC has no clustering. There is sample in every state and almost every county within the United States. Even if county were used, we would likely collapse some of the few small counties without sample with other similar counties with sample rather than try producing the mixed effects models.

In order to evaluate the models, random groups were used to calculate variances of results by running the models over 10 random groups. The sample was not set up to use balanced half sample replication or jackknifing which we felt would have been preferable and more precise. We used only 10 random groups because the smaller sample size of more random groups made it impossible to estimate certain effects due to absence of certain types of cases. The IC conducted in 2000 allows for the use of balanced half sample or jackknife variance estimates. Each has a larger sample and the ability to estimate more of the desired effects.

Composite Estimation

Ghosh and Rao, 1994 describe composite estimation as a natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator. Such estimators can be written as $Y_c = w_i \hat{Y}_{1i} + (1 - w_i)\hat{Y}_{2i}$ where Y_{1i} and Y_{2i} are the synthetic and direct estimators respectively.

One obvious choice of the synthetic estimator is to choose the estimate for a larger domain. For example, if one is making an estimate for a city within a region, then the regional estimator would become the synthetic estimator. This was used in the Consumer Price Index, Cohen and Sommers, 1984. The IC offers a similar choice for its estimates. The optimal value of w_i is:

$$wi = \frac{\left(\text{var}(\hat{Y}2i) - \text{cov}(\hat{Y}1i, \hat{Y}2i)\right)}{\left(\text{var}(\hat{Y}1i) + \text{var}(\hat{Y}2i) - 2\text{cov}(\hat{Y}1i, \hat{Y}2i) + \text{bias}^2\right)}$$

As one can see, if the variance of the direct estimator is very large, more weight is put onto the synthetic estimator. If the bias or the variance of the synthetic estimator is large, then less weight is put on the synthetic estimator.

A problem is that weight values need to be estimated. Thus the actual estimator used will likely be less than optimal. However, with a w that is reasonably close to optimal, the quality of the estimators does not decrease markedly. However, it is also true that reasonable care needs to be given to the estimation of the w.

Numerous authors have tried to address the problem. For instance, using the same weight across all estimators has been proposed ,(Purcell and Kish, 1979), as well as weights based upon the relative sample sizes,(Särndal and Hidiroglou, 1989).

One can estimate both parts of the numerator directly using standard methods. The denominator can be estimated by the difference of the two estimators squared. This estimate is very unstable. As a result, we estimated the bias of the synthetic estimator as

$$bias^2 = \left(Y_{2i} - Y_{1i}\right)^2 - \left(Var(Y_{2i}) + Var(Y_{1i}) - 2Cov(Y_{2i}, Y_{1i})\right),$$

if the value is positive, otherwise the bias is assumed to be zero. This estimate probably makes the weight for the synthetic estimator too large on the average. However, given the size of the variance of the direct estimator, in most of the cases it was impossible to say that the mean of the differences between the two estimates was not equal to zero, so setting most of the biases to zero or near zero did not seem unreasonable. The resulting estimates appeared very reasonable given our general knowledge of how values are affected by areas, industry and size of firm. In the future it is hoped that the weighting issue can be explored further. However, given the current status of the project, we felt this method gave suitable results that could be compared with the models.

Results

Results obtained were very encouraging. Some work must be done to improve the model results. Table A shows the average root mean squared error (rmse) for the sample estimators, two model estimators and three composite estimators. Results shown are for average single premiums, which are similar to those for other variables. Thus, we only show one set for descriptive purposes.

Shown in Table A are the standard errors for the sample estimator and for the model estimator. There are two values for the model estimator. The first assumes the model is unbiased and the rmse for the model is its standard error. The second assumes the model estimates are biased. Bias for each estimator was produced using the method used in making composite estimators and included with the standard error to produce rmse. This was done because the results indicate it is likely that at least some of the model estimates are biased at this time.

The final three composite estimators are (1) the composite of the sample and model assuming the model has no bias, (2) the composite of the sample and model assuming the model is biased and (3) the composite of the sample and estimates for a larger cell. For instance, the composite estimate for a cell determined by a state and size class could be composited with the estimate for the same state or the same size class. Estimates for cases with a single cell identifier, such as state, were combined with the national estimates.

Several things should be noted about these results.

- For the national estimates and estimates for groups defined by only one cell identifier, such as state, the model estimates have about the same errors as the sample under either the biased or unbiased assumption.
- However, for multiple item defined cells, the model estimates do not deteriorate as quickly as the sample estimate, but under the assumption of non zero bias the model deteriorates faster than just the standard error of the model.
- The composites all improve the estimates, especially the composite which assumes the model is unbiased. The composite of the sample estimate with a larger cell estimate appears to be better than the estimate obtained when assuming the model is biased. The gains of these three estimates over the sample increase as the cell sizes for the sample become smaller.

Table B gives the mean deviation (MD)between sample and model results by groups of estimates. Table B also shows mean absolute deviations of the sample(MAD) between sample and the model and the sample and the 3 composite estimators. This table shows that the average deviation over all but one group, size/state, between the sample and model is positive. This is the indication that the model is biased. One can also see that the mean average deviations follow the size of the rmse's in the previous table. However, the composite estimators where bias was assumed to be non zero, tend to have slightly less distance between them and the sample. This is caused by the assumption of bias which pulls the composite towards the sample estimate more than when no bias is assumed. When no bias is assumed, if the sample estimate deteriorates the sample estimate is

ignored and the model estimate dominates the results and there is less shrinkage of the composite toward the sample.

Table C shows the standard deviation of the estimates of average single premium within the groups for the 5 estimators. In the past many small group estimates were basically giving a mean for all estimates which, while possibly having some good statistical properties, really gave the user no useful information. The standard deviation of the groups of estimates should reflect two values, the between group variation in expected values and the size of the errors of the estimates. Thus, if one looks under State on Table A, one notes that for the sample estimate the average error is 72 while on Table C, the corresponding value is 173 which can be thought of as an estimate of average error from both sources. Thus, one could estimate the between expected values deviation as the square root of the difference in the squares of 173 and 72. The value is 157. A similar value for the composite of the sample estimate with that of a larger cell for the State estimates is 144. This slightly smaller value probably reflects the shrinkage of the estimates towards the means of the large group estimates. As one can also see, all the groups have sizable standard deviations which increase with the average error
for the group. This reflects the increasing variation of the larger groups of cells and the increased errors of the estimates.

In any case the estimates of the between cell expected values is rather large, reflecting a range of expected values of about 600 on a set of values with an overall mean of slightly more than 2000. This range of values and visual checks indicated generally very reasonable sets of estimates for the cells.

Conclusions

The estimates produced for smaller cells using either the compound predictions for the entire frame or the compositing to larger cells both appear to have promise to greatly improve the rmse 's of the small cell estimates. Both produced reasonable values that seem to provide estimates which mirror the variation in expected value of the sample. As such all the new estimators appear to have promise.

Of special importance are the estimates produced using compound model predictions for the frame units. Although it appears that the estimates produced during this trial had some slight bias which increase the rmse's, it is hoped that the addition of perhaps one or two variables, such as, county, which were not available for this task, could lessen this bias or eliminate it. Given the size of the estimates of the average standard deviation for the model estimates, relative to those of the sample estimator, combined with the fact that these errors do not greatly increase as the size of the cells decrease, could result in estimates which are a vast improvement over the sample estimates for smaller cells

In the future, work will proceed by trying models with a slight increase in the numbers of independent variables. The array of dependent variables will also be increased. Also intended are use of better replicate variance estimation methods which will included half sample methods and a larger number of replicates. Finally, work may be extended to certain small cell estimates from other AHRQ surveys. This will be done where there are variables which have similar structures where the compound modeling process should apply, that is, variables with a large chance of zero, with a conditional distribution of non zero values to which a standard small area model can be applied, for the units which have positive values.

Table A

Group Average Estimated Root Mean Squared Errors-Single Premiums						
	Sample	Model/Unbiased (Std Only)	Model/Biased (MSE)	Composite Sample-Model/Biased	Composite Sample-Model/Biased	Composite Sample Group/Larger Group
National	20	23	24	20	20	NA
State	72	72	79	64	67	44
Industry	51	52	55	48	49	42
Industry/State	267	89	143	80	104	94
Size	29	27	28	25	25	23
Size/State	140	75	131	64	85	84
Industry/Size	119	66	99	60	74	62
Industry/Size/State	482	95	210	90	150	140

Table B

Group Average Estimated Mean Differences and Absolute Differences - Single Premiums					
	Sample/ Model (MD)	Sample/Model (MAD)	Sample/ Composite of Sample and Unbiased Model(MAD)	Sample Composite of Sample and Biased Model (MAD)	Sample Composite of Sample Group Group and Larger Group (MAD)
National	12	12	4	2	NA
State	15	30	36	14	20
Industry	9	17	17	10	23
Industry/ State	16	191	175	125	136
Size	8	9	6	3	5
Size / State	-24	138	122	63	55
Industry / Size	25	93	82	47	108
Industry/ Size /State	20	352	337	239	243

Table C

Standard Deviations among Estimates within Groups - Single Premiums					
	Sample	Model	Composite of Sample and Unbiased Model(MAD)	Composite of Sample and Biased Model (MAD)	Composite of Sample Group Group and Larger Sample Group (MAD)
State	173	176	169	172	151
Industry	125	130	116	116	98
Industry/ State	329	198	196	241	229
Size	85	91	93	88	79
Size / State	260	192	192	224	228
Industry / Size	218	155	150	168	146
Industry/ Size /State	582	219	221	389	375

References

Agency for Healthcare Research and Quality, Tables of results from the Medical Expenditure Panel Survey, Insurance Component. 2001, August 28. Available from URL: http://www.ahrq.gov/Data_Pub/IC_Tables.htm.

Battese GE, Harter RM and Fuller WA (1988). An error-component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* 83, 28-36.

Cohen M P and Sommers JP (1984). Evaluation of the Methods of Composite Estimation of Cost Weights for the CPI. *Proceedings of the Section on Business and Economic Statistics*. American Statistical Association.

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

Purcell NJ and Kish L. (1989). Estimation for small domain. *Biometrics*, 35 23-30.

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.

Särndal CE, and Hidiroglou, MA (1989). Small domain estimation: a conditional analysis. *J. Amer. Statist. Assoc.* 84, 266-275.