COMPARISON OF DIRECT, MIXED MODEL, AND BAYESIAN METROPOLITAN STATISTICAL AREA ESTIMATES FOR THE INSURANCE COMPONENT OF THE MEDICAL EXPENDITURE PANEL SURVEY (MEPS)[1]

Robert M. Baskin[2] and John P. Sommers[2],
[2]Agency for Healthcare Research and Quality (AHRQ)
540 Gaither Road, Rockville, MD 20850

Abstract
The Medical Expenditure Panel Survey is a family of sample surveys. The Insurance Component of MEPS provides national and state level estimates of insurance offered and provided by employers in the U.S. In recent years the demand for reliable data at the state level and below, regarding healthcare insurance has greatly increased. Previous research has been conducted to produce direct design-based estimates using the MEPS Insurance Component design structure. However, the number of Metropolitan Statistical Areas (MSAs) for which direct estimates can be produced with acceptable reliability is limited. In this paper, we evaluate mixed models and Bayesian models, incorporating a time covariate, to produce MSA level estimates for smaller MSAs in ten states. We examine estimates of MSE and RSE of two types of estimates based on direct and indirect estimation techniques.

KEY WORDS: complex sample, design bias, mean squared error

1. Introduction

The Medical Expenditure Panel Survey (MEPS) is a family of sample surveys conducted by the Agency for Healthcare Research and Quality (AHRQ). The Insurance Component of MEPS (IC) is conducted by the U.S. Census Bureau for AHRQ. The IC is a stratified probability proportional to size sample of private and public sector employers designed to collect data on the number and types of private health insurance plans offered, benefits associated with these plans, premiums, contributions by employers and employees, eligibility requirements, and employer characteristics. Currently, AHRQ publishes extensive tables of estimates at the level of nation, state and top twenty Metropolitan Statistical Areas (MSAs).

The purpose of this study was to evaluate the use of Small Area Estimation (SAE) techniques for making MSA estimates for small geographic areas that are currently unpublished. The method of evaluation is to estimate the design mean squared error (MSE) for each of the estimators evaluated. Section two will lay out the IC sample design. Section three will discuss the techniques used for estimation and the method of estimating the MSE of each. Section four will give the results with some discussion and the last section will give conclusions and further possible research.

2. The Sample Design

As was mentioned, the IC sample design is a stratified probability proportional to size sample of private and public sector employers. The frame is the List Sample, a nationally representative sample of employers developed from Census Bureau list frames. All of the tables posted on the MEPS-IC Web site are derived from the List Sample.

The private sector frame is stratified by size of establishment and by industry classification whereas the public sector frame is stratified by size of the public sector entity. Since 2002, the sample has been allocated to the two frames in a way that allows publication of state level estimates for fifty states plus the District of Columbia. It should be pointed out that the sample size at the MSA level is a random number. However, systematic sampling is used within the strata and the strata are sorted by geographic area so the sample size at the MSA should be close to proportional to the MSA size.

Selected units are mailed a questionnaire which asks about health insurance plan characteristics, and firm-level (company) characteristics. The establishment-level characteristics include information such as number of active employees , whether or not establishment offers health insurance, number of plans offered, number of employees eligible

[1] The views expressed in this paper are those of the authors 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.

for health insurance and number enrolled (full-time and part-time employees separately), and workforce characteristics (% women, union, over 50 years old, by wage level). Plan information includes premiums, contributions, plan types, self-Insured or fully-Insured, enrollments, deductibles and co-payments, some plan benefits, and "Fringe" benefits. Information on firms includes size of firm, industry, age of firm, retiree offerings, and employee characteristics. Details of the MEPS IC sample design have been previously published in Sommers (1996).

3. Estimation and MSE Methods

Direct estimates are currently produced for the nation, for states, and for the largest twenty metro areas. These estimates have standard errors (SEs) estimated in production by the method of random groups. For current levels of published estimates the relative standard errors (RSEs) are acceptable and the national level RSEs are generally below 5%. However, the RSE of the direct estimates for many metro areas smaller than the top twenty may be larger than 30% which is generally considered unacceptable for publication.

In this paper we investigate estimation of mean premiums and mean employee contributions for smaller metro areas within ten selected states using mixed model and Bayesian model estimates which borrow strength across time and across domains. MSE estimates were also produced using ten replicates within each state. These estimates were produced for the years 2000 through 2005.

3.1 Data Used

Because not all MSA level estimates are run in regular production, special runs were required to produce the data used in this study. Ten states, including New York, Vermont, Louisiana, and South Dakota, were selected to represent various types of states in production. Within these states, estimates for three to five of the MSAs, excluding the largest MSA in the state, were produced for mean premiums and mean employee contributions. The excluded MSA as well as the rest of the observations in the state were then grouped together to form a 'rest of the state' estimate for each of the ten states. For each of the ten states, ten Jackknife replicates were created and used to produce point estimates for each replicate as well as estimates of SEs.

3.2 Model Estimates

Model estimates based on a composite type estimate, frequentists mixed model estimates, and Bayesian model estimates are used to compare to the direct estimates. The MSEs of these model estimates are calculated and compared to the RSEs of the direct estimates. The composite is a composite between the MSA direct estimate and the rest of the state but does not borrow strength over time. Both the mixed effect models and the Bayesian models borrow strength over domains as well as over time.

There were three time models that were investigated in the process of modelling the MSA estimates. First, a simple linear model over time was used. This actually proved to be the best approach in the long run. A second time model was to incorporate higher order time effects. The inclusion of higher order terms increased the estimated MSE of the estimates so this approach was abandoned and the results are not shown. An auto-regressive (AR) time model was also attempted. This approach also increased the MSE. However, it may be possible to improve the AR through the covariance term so the AR approach will be considered for future research.

There are two models considered for borrowing strength over geographic domains. In the simpler model the MSA estimate for the year is modeled as a random effect around the 'rest of the state' estimate. The 'rest of the state' estimates typically had ten times as much sample as any of the MSAs investigated and thus were much more stable. A more elaborate version of this model included a term for a random effect for each of the ten 'rest of the state' estimates as random around the national estimate. This model produced smaller MSEs so this is the model discussed below.

3.3 MSE Estimates

The direct estimate is design unbiased so the RSE of the direct estimate is equivalent to the MSE of the model based estimates. However, because the model based estimates are in general design biased, an approximation of the bias of the estimate is needed in order to calculate the MSEs of these model estimates. Estimation of the design bias of a model based estimate is difficult. One method that has been used since the 1970s, referenced in Rao (2003), is to take the difference of a design biased estimate and a design unbiased estimate and use this difference as the estimate of bias. It is true that this estimate of bias has the correct expected value but it is as highly variable as the design unbiased estimate that is used in the process. However, since this seems to be the only way to proceed to get an estimate of design bias at this point in time, this is the basic approach that is used in this paper. One small modification was available for IC data because of the estimation at the replicate level. For each replicate the difference between the direct replicate estimate and the model based replicate estimate is a valid estimate of design bias. Thus, the unsigned average of these replicate differences is an estimate of the design bias of the model based estimate. It is important to note that this puts the model based estimate at a distinct disadvantage because the weakness of the direct estimate, the high variability, contributes to the bias estimate of the model based estimate. Therefore, under these circumstances, if a model based estimate can be shown to have smaller MSE than the direct estimates' RSE, it can be interpreted as the model based estimate being profoundly better than the direct estimate.

4. Results

4.1 Direct MSA Estimates

The estimates under consideration are mean of premiums and mean of employee contributions. These estimates are produced for 33 small MSAs in ten states along with an estimate for the 'rest of the state' for each of those states. These estimates are for the years 2000 to 2005.

As can be seen in Table 1 the MSAs can exhibit changes over time that appear to be contrary to the state and nation. For example, Huntsville Alabama had four consecutive years, from 2002 to 2004 of decreasing employee contributions for health insurance even though the state and nation showed an increasing trend over that time period. This type of MSA behavior is considered to be a problem of small sample size since there is no evidence that insurance rates were actually decreasing in Huntsville Alabama in that time period, contrary to the behavior in the rest of the nation. One goal of using the model based estimation was to see if this behavior could be eliminated while still reducing the overall MSE.

4.2 MSE estimates

The average MSEs for the four types of estimates are given in Table 2. The first column contains the RSE of the direct estimate. The average RSE for the direct estimate of small MSAs in the years 2000 to 2005 is often above 30%, which as a rule of thumb is considered unpublishable. The next three columns contain the relative MSE (RMSE) of the three model based estimates. The composite estimate has RMSE that is approximately half the RSE of the direct estimate. The composite estimate, even with the bias included, is on average twice as accurate as the direct estimate. However, the composite does nothing to attempt to control the type of time changing behavior seen in Huntsville Alabama. The model estimates that attempt to borrow strength across time, namely the mixed model estimates and the Bayesian estimates, have smaller RMSE than the direct but not as small as the composite estimate.

There are considerations about the technical advantages of the model based estimates. The composite certainly has the smallest RMSE and takes a form that is easy to estimate in production. However, it does not account for the time varying nature of the data. Both the mixed model estimates and the Bayesian estimates used a linear time model and had similar RMSE. The mixed model estimates were produced using Proc Mixed in SAS which means that they could be produced in production without using new software. The Bayesian estimates theoretically have good properties, have slightly smaller RMSE than the mixed model estimates and thus have smaller error rates than the direct estimates. But Bayesian estimates were produced in R using the package BRugs, and thus would require a change in processing if they were to be used in production, or a separate process altogether.

5. Discussion

The MEPS IC sample design allows estimation of national, state, and some metro area estimates. This work investigated the use of model based estimates for small MSAs in the IC. The model based estimates had different advantages but each improved the overall error rate of the small MSA estimates compared to the direct estimate. The composite estimate had the smallest overall error but did not incorporate a time component. The mixed model estimates had a time component, improved the overall error rate, and were slightly easier to produce than the Bayesian estimates. The Bayesian estimates incorporated a time component, had slightly smaller overall error rate than the mixed model estimates, as well as the direct estimates, and have good theoretical properties. The investigated model based estimates improved the error rate to the point that they should be considered for publication. There is still room for improvement in the models under consideration. If this approach of MSA level estimation is pursued, then an investigation into the sample sizes for which these models will produce valid estimates is necessary.

References

Rao, JNK. Small Area Estimation, 2003, John Wiley and Sons

Sommers, JP. List Sample Design of the 1996 Medical Expenditure Panel Survey Insurance Component. Rockville (MD): Agency for Healthcare Research and Quality; 1999. MEPS Methodology Report No. 6. AHRQ Pub. No. 99-0037.

Table 1. Direct Estimates of Average Employee Contributions for Small MSAs in Alabama

Average Employee Contributions			
	Huntsville (N)	**Mobile (N)**	**Rest of State (N)**
2001	**$600.14 (47)**	**$541.67 (38)**	**$516.40 (395)**
2002	**$793.93 (35)**	**$472.11 (50)**	**$641.63 (340)**
2003	**$657.15 (58)**	**$475.21 (53)**	**$632.64 (360)**
2004	**$530.21 (50)**	**$753.97 (55)**	**$630.55 (407)**
2005	**$441.61 (42)**	**$503.72 (26)**	**$783.25 (396)**
2006	**$877.47 (39)**	**$1255.59 (38)**	**$806.11 (395)**

Table 2. Average Error Estimates for Average Employee Contributions

Average Error Estimates				
	Direct RSE	**Composite RMSE**	**Mixed Model RMSE**	**Bayesian Estimate RMSE**
2000	**0.39**	**0.18**	**0.29**	**0.29**
2001	**0.31**	**0.21**	**0.30**	**0.25**
2002	**0.47**	**0.20**	**0.20**	**0.22**
2003	**0.25**	**0.13**	**0.23**	**0.14**
2004	**0.34**	**0.13**	**0.16**	**0.14**
2005	**0.24**	**0.14**	**0.18**	**0.15**