Some Relationships and Possible Models for State Estimates from the Medical Expenditure Panel Survey, Household Component

John Sommers

Agency for Healthcare Research and Quality Working Paper No. 06007

November 2006

Suggested citation: Sommers J. Some Relationships and Possible Models for State Estimates from the Medical Expenditure Panel Survey, Household Component. Agency for Healthcare Research and Quality Working Paper No. 06007, November 2006, http://gold.ahrq.gov.

Some Relationships and Possible Models for State Estimates from the Medical Expenditure Panel Survey, Household Component

John Sommers
November, 2006

ABSTRACT

In recent work at the Agency for Healthcare Research and Quality, efforts have focused on production of direct estimates of expenditure variables with data from the Medical Expenditure Panel Survey, Household Component (MEPS – HC). For some more common medical expenditures for larger states this work has been very successful. However, quality direct estimates cannot be made for many important variables for many states. In order to produce better state estimates for many variables, model based small area estimates will be necessary. To produce such estimates, models need to be developed to serve as the basis of these types of estimates.

In this paper relationships among estimates for individual states are examined over time as well as the relationships between estimates for individual states and national estimates over time. We show there are high rank correlations among state estimates over time. Further we show, that most estimates for individual states correlate well with national estimates over time. The former correlations indicate that values for individual states are somewhat fixed in relation to those of other states over time. The latter correlation indicates that change for individual states closely follows the national change over time.

Finally, we briefly outline how this knowledge might be useful in the development of a Bayesian small area estimation process and propose work proceed in this direction in the efforts to develop better state level estimates with the MEPS - HC data.

John Sommers
Mathematical Statistician
Center for Financing, Access and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: John.Sommers@ahrq.hhs.gov

Background

In recent years work has taken place to expand knowledge useful in the production of State level estimates using data from the Medical Expenditure Panel Survey, (MEPS) Household Component (HC). This work (Sommers, 2005) has shown that one year annual direct estimates with reasonable relative standard errors could be made for some classes of expenditures for the variables: conditional mean, type of expenditure, and percent with that type of expenditure. Other work, which has not been published, has shown that for certain types of expenditures, direct estimates can be made for the percent of expenditures paid by private insurance, the percent of expenditures paid out-of-pocket, the percent of expenditures paid by Medicaid, and the percent of expenditures paid by Medicare. The original work also showed that using a simple type of small area estimation technique, estimates with errors smaller than the direct estimates could be produced (Sommers, 2005). However, for some less common types of expenditures., even using the improved small area estimates, the average relative mean squared errors (rmse's) for smaller states were between 10 and 20 percent on average and the maximum rmse's were over 30%. Thus, there was still need to find methods to further improve state level estimates.

A second study focused upon the potential effect of using more than one year of data from the HC (Sommers, 2006). Use of data from 3 years has been a method used by the Census Bureau to produce estimates of income (Census Bureau Website a, 2006). Larger sample size cuts the error. However, the gain using three years of data from the MEPS-HC is not a much as one would hope for. This is because the HC uses panels of sample units over two years of data collection (Cohen, 2000). The same sample in two years of data causes correlation in the data across years. This correlation limits the reduction in error from the increased sample size to a value less that would be obtained using data from across years which was independent. Sommers, 2006, showed that on average using three years of data resulted in about a 35% reduction in relative standard errors for a large group of direct state estimates made for a variety of expenditures for several health conditions, such as, obesity and diabetes. Although, this was only a small portion of the possible state estimates that could be made, the average reduction probably represents a good 'rule of thumb' for what can be gained using three years of MEPS-HC data.

As with estimates using single years of data, one can apply small area techniques to the three year estimates. This further lowers the rmse's compared to those of the direct three year estimates. Even with these results there are many estimates for less common events, such as, estimates for hospital visits, estimates for persons with less common conditions and estimates for smaller states which still have unacceptably large rmse's.

Also these estimates are produced using the expected values of three years of results. Multiple years of data may be difficult to interpret. We don't really know how useful such results are. For instance, such a result would be of little or no value when trying to spot changes or trends. It could take several years to develop a significant difference across three years of results, where two years overlap, because results are highly correlated. Two consecutive sets of three years from the MEPS-HC share two full years of data and common panels of data between the most recent yea. This could make it difficult to find significant changes across sets of years. It would likely be as difficult to find a significant difference between two sets of consecutive 3 year estimates with two overlapping years, as it is to find a difference in estimates using two consecutive single years of data. While the three year data has smaller errors, the change between two sets of three years is heavily dampened by the overlap of data. Thus, although the errors are small the changes in results are also likely smaller than what one would see with single years of data. Smaller errors applied to checking smaller changes may not result in any better information than comparing two single year results which change more but have larger errors.

What is to Come

Presently, we have exhausted the simple options we have to produce state estimates with the MEPS-HC. For some expenditure types, we know direct estimates can be made for states. Since we have succeeded with estimate for some expenditures, it seems that we must simply try direct estimates for groups of states for other variables and possible subsets of the population. For instance, what would the rse's be for estimates of total expenditures for persons with obesity in large states using a single year of data. From sample sizes and comparisons of errors for other estimates with similar sample sizes and structures we can predict the most likely variables and population subsets that will result in direct estimates of reasonable quality using one year of data. One simply needs to try those that might work and find those that do. One can also determine when the simple composite types of small area estimates, already developed, can produce estimates of reasonable quality. Again one must simply apply the methods and decide where they have provided suitable results.

Likewise, one must make the determination if estimates using three years of data are useful.   As with estimates made with a single year of data, this is somewhat of a trial and error exercise. However, for these multi year estimates, further questions must be answered.   For instance, one must also determine if users would find multi year estimates useful.

Some direct estimates for Metropolitan Statistical Areas (MSA's) may also be possible. This could be determined by comparing sample sizes of the MSA's with states for which estimates have already been made. It is likely that similar standard errors could be produced using similar sample sizes for the same types of estimates.

Thus, with an organized, meticulous approach, plus some decisions about use of the simple composite small area estimates and multi year estimates, a good number of quality state and MSA estimates can be produced. However, other methods will still be needed to improve upon these methods and to produce estimates for smaller groups of the population, such as, persons with obesity or diabetes or inpatient hospital visit expenditures.

In order to produce reliable estimates for rarer populations or less common events some type of modeling approach will be required. Modeling approaches might also be used to improve upon some current estimates. This type of approach must be able to take advantage of relationships and patterns among data and results across time and states.

If one could detect patterns across years or states or correlation with other results, one would have more information that could be brought to the estimation than would be available from a single or possibly multiple years of sample for an individual state. The following are some simple hypothetical examples:

- Suppose for the last 10 years the estimate for a particular state had never been statistically different from the national estimate and the state estimates had always had a relative standard error of less than 5%. If in the 11th year the estimate one had an estimate with an rse of 30%, one might feel that a better estimate for the year for the state would be the national estimate or an estimate that moved the sample estimate closer to the national estimate for the year.

- Suppose that the estimates for the past 10 years for a state had never been statistically different from one another. Could there be an estimate produced for the 11th year using the average for the 10 previous years that would better than the direct estimate for the 11th year?

- If one could predict the value for a state using a regression model based upon other easily obtained independent variables, then a second estimate could be made and combined with the direct estimate to improve upon the direct estimate. The new estimator could be obtained using some type of variance component approach, like those examined in Ghosh and Rao, 1994.

Dependent upon the relationships found in the data, all types of possibilities for estimation arise, from time series approaches, such as a Kalman Filter, (Harvey, 1989) or some more complex Bayesian method, such as those outlined in Rao, 2003.

The remainder of this paper discusses some simple relationships that have been found in the data that could be useful to ‘borrow strength and information’ from across states and across years. This information could be used to produce model based estimates which could improve upon the current direct estimates and simple small area estimation methods currently available to produce state estimates with MEPS-HC data.

Searching for Patterns

Ideally, when one searches for patterns within sets of data or estimates, one would like to examine the expected values of those estimates. After all, this is the item one is trying to estimate. However, normally one does not have a set of true values to examine. Sometimes after the fact one might have administrative records that can be used to provide the ‘true’ answer. For instance, if one were making estimates of state sales tax revenues for a year, eventually the state government will have the amount for the year to examine. Medicare payments can eventually be totaled and examined for patterns. Using such information allows one to check for patterns. Which states always have higher than average payments per capita each year? Does the rank in per capita spending stay fixed or vary little from year to year. If such patterns exist then one might be able to use back data along with a current sample to estimate per capita Medicare spending before the final administrative records were available. Use of the patterns and back data could possibly yield a better estimate than the sample alone.

Unfortunately, for most of the items for which state estimates from the MEPS-HC are desired there are no administrative records or truth sets to guide us. All that is available is a series of estimates for each state's results. Dependent upon the quality of the estimates one might be able to discern a pattern among the true state values, if a pattern existed.   However, even if a pattern existed among the expected values, it could be somewhat or totally hidden within the estimates. A simple example can illustrate the problem.

Suppose one was interested in state per capita income and that the true values for each state were always a fixed proportion of the national per capita income value. Thus, for the ith state the value each year would be p(i)*n where n was the national value and p(i) were the proportion for the ith state. If one knew this, a good estimate for an individual state would be to take the average ratio of the past state estimates to the national estimate and multiply by the current national estimate. Such an estimate, allows one to use all the past information plus that from the current year to estimate the state value.

However, even with such a perfect stable pattern of relationships, one might not be able to find the pattern using past estimates. Suppose that for the 50 states and the District of Columbia, if one ranked the values of p(i) that they ranged from .75 to 1.25 and the values took the equally spaced values, .75, .76......1.23, 1.24 and 1.25. If the standard errors for the past estimates were greater than the .01 value that is the relative distance between each states values, then it would be very likely that if one ranked the annual estimates each year for the states, that there would be a positive correlation between the annual ranks, but it would not be the 100% correlation that actually existed between the expected values of the estimates. Furthermore, if the relative errors of the individual state estimates were very much greater than the relative distance between the actual expected values, the annual estimates might show very little correlation, because the ranks of the estimates could shift considerably from the actual ranks of the expected values due to the errors in the estimates.

Given this phenomenon, there could be many patterns within the expected values of the state estimates and few might be found among the sets of state estimates available for the MEPS-HC. For this reason, this examination of past estimates is to obtain hints about what to try.   For instance, if we find certain general patterns among estimates of certain state expenditure types but not others, we might still use the general form of model that the pattern would indicate on other types of expenditure estimates, especially if we find the pattern among estimates for expenditures with relatively low errors, such as, prescription drug expenditures but do not find the pattern among estimates with larger errors, such as, inpatient expenditures. In the final analysis, we will still have to assess the mean squared errors of our estimates using any model in order to determine whether any set of estimates are better than any of the currently available direct or composite estimates.

Rank Correlation of State Estimates over Time

One of the first patterns considered was rank over time of various use and expenditure variables among the 28 states that we could estimate from 1996 to 2003. We decided not to use the year 2000 due to a very much smaller sample of sample units for the year that led to very high standard errors for state estimates compared to other years. Since we had learned from other sources that the rank and level of estimates for many items do not vary radically over the years, there may be similar patterns among HC data. For instance, the state poverty rates are very similar over periods of time. States with high rates stay the highest over periods of time and states with low rates stay low (Census Bureau Website, b)

The types of expenditures analyzed were any, dental, prescription drugs, outpatient, and inpatient. The first set of expenditures analyzed were the percent with an expenditure. Of the 21 pairs of years the rank correlations were significant at the 1 percent level for all 21 combinations for the any expenditure, dental expenditure and outpatient expenditure classes. Prescription drugs were less correlated and had 3 of the 21 pairs that were not significant at the 5% level. The averages for the 21 rank correlations for percent with an expenditure for these four classes were 72.0, 82.3, 56.7 and 75.3 percent respectively. These seem quit impressive given the average relative standard errors of some of the sets of estimates is well over 10 percent and the data are as many as 8 years apart.

The percent that used inpatient facilities, estimates with the highest relative standard errors among the groups, showed some relationships. Given the small range of the estimates, approximately 6 percentage points, and high relative standard errors, many over 20 percent, only eight of the sets of years of inpatient usage rates had rank correlations which would have been significant at the 5 percent level.

The rank correlations of state results for conditional mean expenditures for the five expenditure types for the years from 1996-2003, were far less significant than those for the percent with an expenditure. Perhaps that should be expected given the larger relative standard errors for this type of estimate versus estimates of percent with an expenditure (Sommers, 2005). For instance, for 2003 data the average rse for conditional means for the 5 types of expenditures for the 28 states were all at least 10 percent, except for prescription drugs which was 8.7 percent. The average rse for the same 28 states for 2003 for the percent with an expenditure were on average about half the value for the average rse for estimate of the conditional means for the same type of expenditure. For instance, the average rse for estimates of the percent with a prescription drug expenditures was only 3.4 percent.

For the same set of expenditures and years, the rank correlations for the 28 state estimates were calculated for the percent paid by self and family. These yielded poorer results than those for conditional mean expenditures. This was not surprising given that the average relative standard errors for the percent paid by self and family is higher than that for the conditional mean expenditures. However, as with the conditional mean expenditures dental and prescription drugs showed some good average rank correlations over the sets of years of 36.0 and 49.3 percent respectively.

<u>Correlations with National and Regional Estimates</u>

Although, the rank correlations indicate that the various state estimates are related over time, in order to borrow information and strength across estimates, we must show the state estimates can be predicted by other available variables across the years or correlate over the years with better estimates, such as, national and regional estimates.

We first tried to find other correlates, such as, per capita income, age, etc. which predicted our HC results for usage and expenditures. However, thus far, none have been found that predict well enough to improve upon the sample estimates.

We also looked at the correlations between each of the state estimates over the period with its regional estimate and the national estimate for the same year. The results from this process were of some interest. The results seem to show that there are strong relationships among many of the expected values. Table 1 is helpful in this examination.

TABLE 1

Average Percentage Correlations of Annual National and Regional Estimates with State Estimates of Conditional Mean Expenditures for 3 Types of Expenditures, Years 1996-2003 for the 3 Groups of States Ranked by Size

		Expenditure Types		
Correlate	**State Size Group**	**Any**	**Prescription Drug**	**Inpatient**
National Estimate				
	1	79.7	97.1	47.7
	2	79.7	91.2	32.6
	3	52.9	93.7	3.2
Regional Estimate				
	1	61.1	83.6	42.7
	2	59.3	77.7	34.5
	3	36.5	81.7	6.0

To produce this table we took the 28 states being considered and divided them into 3 groups of 10, 9 and 9 states respectively. The states in the first group were the 10 largest, the next 9 the next 9 largest and the last 9 the 9 smallest states among the 28. We then calculated the Pearson correlations of each state's set of conditional mean expenditure estimates with the corresponding national and regional estimates for the 8 years. These correlations were then averaged for the set of states in each size group. The results shown in Table 1 are illustrative of the patterns found. Among these patterns are:

- The averages for the last group, which has the worst estimates are the lowest. This was expected because the level of error is much greater and masks the actual expected values.

- The averages are best for estimates for the groups any expenditures, dental, prescription drugs and office based care. They are worst for inpatient and outpatient care. This again is probably due to the higher rse's of the latter two types of estimates.

- The correlations are lower with the regional estimates, but follow the same patterns as those with the national estimates. Again this is possibly because of the higher errors for the regional estimates than the national estimates which may obscure correlations in expected values.

- The correlations follow the same patterns for the variables, percent of persons with an expenditure and percent paid out of pocket. However, they are almost always lower than the corresponding correlations for the conditional mean expenditures, although the values are still significant for the groups of larger states and the types of expenditures, such as, dental and prescription drugs, with lower relative standard errors.

Observations and Possible Future Work

As was discussed earlier, some success has been had in the production of state estimates using direct estimates. Improvements were made to these using a composite small area estimation technique which borrows strength and stability from regional estimates. However, many of the results although they have smaller rmse's than the direct estimates, are still not satisfactory, especially for smaller states and less common expenditures.

In the past, the author has attempted to find auxiliary information which could be used to produce prediction models which would use results from all states to produce estimates through model predictions for each state. No success has been had producing regression models of adequate quality that use state characteristics, such as, per capita income, to make model predictions for the various types of state estimates, such as, conditional mean expenditures. If such models were possible, small area techniques, such as, Empirical Bayes methods or Fay-Herriot type models could be tried. These approaches tend to weight the two estimates for the same quantity to produce a third estimate (Rao, 2003). If there were a model estimate the small area estimates would be a weighted combination of the model and direct estimates. Without additional information these estimates use the national or regional mean as the alternate estimate and the new estimate developed is very similar to the composite estimate already produced (Sommers, 2005). Without new information these approaches give similar small area estimates to the those already developed and would likely not show a any significant improvement.

The correlations discussed in this paper are new and potentially important information. Rather than produce second estimates for each state using models based upon the economic, demographic and health statistics for a state, these correlations relate the state estimates to similar estimates for other states, regional estimates, national estimates and the estimates for the same quantity for the same state for other years.

This leads to the possibility of some type of time series modeling which might allow one to produce an estimate based upon a prediction made by combining information from past state and national estimates with the current estimates. Thus, strength could be 'borrowed' over time and relationships with the national estimate, rather than through common characteristics.

An example of a case where such information might lead to an improvement in a state estimate is the results for the percent of persons who had a dental expense in the state of Arizona for the year 2000. National estimates for the percent of persons with a dental expense were approximately 40 percent from 1996 through 1999. For the same years the average for the state of Arizona for the percent of person with a dental expense was about 32 percent and the variation was relatively small over the 4 years. The 2000 national estimate for this percentage was slightly less than 40 percent, a slight drop from the average. Given the pattern of the previous years a model would likely predict the Arizona result for the year 2000 as something less than 32 percent. The direct estimate for the year 2000 for Arizona was barely 20 percent, but this estimate had a relative standard error of over 30 percent, a very poor estimate and much worse than those for the previous 4 years. Given this information, most people and most small area estimates based upon a model which adequately reflected the correlations with the national estimates and the previous years' estimates, would probably produce an estimate of perhaps 30 percent for Arizona for the year 2000. This is likely a vast improvement upon the direct estimate of 20 percent. One would have further evidence that this was an improvement upon the direct estimate, when one checked the direct estimates for the percent of persons with a dental expense in Arizona for the years 2001 through 2003. These estimates returned to the pattern of being over 30 percent for all three years.

In order to take advantage of this information, the next work in our research will be to incorporate this information into models and produce predictions based upon this model using the individual state and national estimates for a series of years.

Because we have the BUGS (Bayesian inference Using Gibbs Sampling) software (Spiegelhalter, et. al,, 1997), readily available, which allows us to produce conditional estimates using Gibbs sampling methodology (Brooks, 1988) we intend to first assess estimates developed using Hierarchical Bayes (HB) methods(Ghosh and Rao, 1994). This method produces estimates by assuming the parameter of interest has a distribution. The distribution of the direct estimate is conditional upon the parameter. One then solves for the conditional expectation of the parameter given the estimate under the distributional assumptions. This is the Hierarchical Bayes estimate for the parameter.

A simple example of the Hierarchical Bayes method using a model over time follows: suppose we assumed that the expected value of a direct state estimate for any year was the same every year. Thus, for the ith year $E(s_i) = S$ for all years.

Given this value of S, the distribution of s_i is $F(s_i|S)$.

Further, assume the distribution of S is G(S).

The joint distribution of s_i and S is $F(s_i|S)*G(S)$ for each year.

If there were n years of estimates and each estimate s_i were independent of the others then the joint distribution of all the n years of estimates given the value of S would be

$$\left[\prod_{1}^{n} F(s_i|S)\right].$$

The joint distribution of the s_i's and S would be

$$\left[\prod_{1}^{n} F(s_i|S)\right]*G(S).$$

Using these assumptions, the estimate for the year would be $E(S|s_1, s_2 .....s_n)$.

This is a very simple example, but it shows that basically by putting a distribution on the parameter one can then work through the problem to find the expected value of the parameter given the sample estimates. In all but the simplest cases this requires an approximation be made which is the purpose of the BUGS software and Gibbs sampling.

To produce HB estimates for states, using the HC will require a more complex model setup. A simple possible setup which would reflect the correlation with the national estimate and the high rank correlation for a state would be the following:

Let s_{it} be the estimate for the ith state for the year t.

Assume $E[s_{it}] = \mu_t + S_i + e_{it}$ for each of the years where $E[S_i] = E[e_{it}] = 0$ for each i and t.

The mean of the national estimate for the year t is assumed to be μ_t. The distribution of μ_t would have mean that could vary from year to year, perhaps with a time trend. Given this set up the means for each state on the average are the mean of the national for the year plus a value S_i which is the same for every year. Thus, each state estimate would have a mean for each year which moves in unison with the national mean is on the average is the same distance from the national mean each year. This would create a correlation between state and national estimates across years. Further, since the distance from the national mean for each state is approximately the same each year, then the ranks for the state means have a high rank correlation.. States with a large value of S_i would have a large expected value each year and a state with a small value of S_i would tend to have small expected value each year. Thus, the model would give the estimates the two properties over time that were found in our data examination, rank correlation and correlation between the state and national estimates.

A model of this nature will be the first tried in the quest to produce HB estimates for states. The data from the year 1996 through 2004 will be used to produce conditional means for each state for 2004 given the series of estimates for all states and the nation.

The BUGS software will be used to calculate the estimates for the conditional expected values for the states.

Because of the complexity of calculating rmse's for such models, the author intends to calculate the mean squared errors of the estimates using half sample replication (Wolter, 1985). In this case this means that the BUGS software will be used to compute multiple estimates of the conditional expected values. These values and half sample estimates of the direct estimates can be used to calculate the mean squared errors, incorporating variance and bias, of the model estimates. These details will be included in future papers describing the production of the model estimates.

References

Brooks, SP (1988). Markow Chain Monte Carlo Method and its Applications, *Statistician*, Vol. 47, 69-100.

Census Bureau. Web site a: http://www.census.gov/hhes/www/income/incomestats.html

Census Bureau. Web site b: http://www.census.gov/hhes/www/housing/ahs/statedata.html

Cohen SB. Sample Design of the 1997 MEPS_HC. Rockville (MD); AHRQ: 2000. Meps Methodology Report No 11 AHRQ Pub No 01-0001.

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

Harvey, AC (1989). *Forecasting Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.

Rao, JNK (2003). *Small Area Estimation*, Hoboken: Wiley.

Sommers J. *An Examination of State Estimates Using Multiple Years of Data from the Medical Expenditure Panel Survey, Household Component.* Agency for Healthcare Research and Quality Working Paper No. 06004, May 2006, http://gold.ahrq.gov.

Sommers, J. P. *Producing State Estimates with the Medical Expenditure Panel Survey, Household Component.* Methodology Report No. 16. December 2005. Agency for Healthcare Research and Quality, Rockville, Md. http://meps.ahrq.gov/mepsweb/data_files/publications/mr16/mr16.pdf

Spiegelhalter, DJ, Thomas, A, Best, NG and Gilks, WR (1987, BUGS: Bayesian Inference Using Gibbs Sampling, Version 6.0, Cambridge: Medical Research Council Biostatistics Unit.