VARIANCES FOR MODELS USING 'AGED' DATA

John Paul Sommers, Agency for Health Care Policy and Research
Executive Office Center, 2101 E. Jefferson St., Rockville, MD 20852

KEY WORDS: Microsimulation, Variances

Background

In order to inform policymakers about the economic consequences of a particular policy change, microsimulation models are often used.

In order to make economic estimates for many national totals for a given population subset for a specific year, economic microsimulation models must make use of estimates derived by combining results from several different sample surveys [Citro and Hanushek, 1991]. This requirement to use multiple surveys arises from two sources.

1. Many surveys are not conducted every year. For instance, medical care expenditure data was last collected in the National Medical Expenditure Survey of 1987 (NMES, sponsored by the Agency for Health Care Policy and Research (AHCPR)). Many current estimates of medical expenditures rely on NMES results adjusted with results from other sources.

The year of data collection is often prior to the year of interest. In this case, projection to a future year (e.g. 1992) may involve the combination of data from several surveys with assumptions about changes in demographics to produce an estimate. For instance, to produce a future estimate of medical expenditures, one might develop trends in changes in usage from the National Health Interview Survey (NHIS), sponsored by the National Center for Health Statistics (NCHS), apply this change to an expenditure per capita developed from NMES, adjust for trends in population growth from the Census and adjust for projected inflation.

2. Certain data elements are not collected in all surveys. For instance, if NHIS collected information on demographic characteristics and health status to represent the entire nation and a researcher independently developed a regression model relating health expenses to these characteristics, using data from another source, then one could estimate health expenses for each sample person in the HIS and use the weighted sum of these estimates to estimate national health expenditures.

Such techniques are regularly be used to make estimates of current or future economic totals for the nation. However, in spite of this important usage, little has been done to measure the error associated with such estimates [Citro, and Hanushek, 1991].

As with any statistical estimator, the estimator can have variance or random error and bias or fixed error. Variance can be caused by the random sampling process and other random errors caused by interviewers, respondents and processing. Bias can be caused by systematic errors, such as, a tendency by all respondents to overestimate a particular expenditure. Bias in microsimulation can also be caused by an erroneous assumption, such as, assuming a certain reaction by consumers to a tax. If consumers do not react in that manner, estimates could be biased.

Such bias can only be determined by comparison with "truth." When comparing estimates of multiple scenarios, which vary with underlying assumptions, an estimate of the size of sampling error can be useful. For instance, when comparing estimates of projections of cost under alternate scenarioes with current costs, it is of great benefit to the policymaker to realize that differences between results obtained under the various scenarios or from current values are not statistically significant when one considers the size of sampling errors for the estimates.

Estimators and Variances

'Aging' is the process of reweighting data from a base survey conducted in a previous period, using data from more recent surveys to provide an updated estimate. In the case we will examine, data from the 1977 National Medical Care Expenditure Survey (NMCES) will be adjusted to the year 1987 using the Current Population Survey (CPS), the 1977 and 1987 National Health Interview Surveys (NHIS) and the Consumer Price Index (CPI). The process of 'aging' data is implemented in the following manner.

Let wgt_{ik} be the weight for the kth unit within the ith population cell defined by demographic factors and exp_{ik}, the health care expense value for the unit from the base survey, NMCES. If for each unit we define a new weight nw_{ik} and new expense $nexp_{ik}$ where

$$nw_{ik} = wgt_{ik}\bullet\frac{P_i}{O_i} \text{ and}$$

$$nexp_{ik} = \exp_{ik}\bullet CP\bullet CU_i \text{ where}$$

P_i is the cell population from the CPS (1987)

O_i is the cell population from the NMCES (1977)

CP is a price change measured by the Consumer Price Index (CPI) from 1977 to 1987 and

CU_i is a change in utilization measured by the ratio of the per capita visits for a given health care service from the 1987 NHIS relative to 1977 NHIS for the ith cell.

An estimate of expenditures across a set of demographic cells $i \in S$, which are mutually exclusive and exhaustive is

$$E = \sum_i\sum_k nw_{ik} \bullet nexp_{ik}$$

$$= \sum_i P_i \bullet PC_i \bullet CP \bullet CU_i \text{ ,}$$ where the sums are for all subcells i in S and PC_i is the per capita expenditure for the ith cell from the NMCES.

Because the factors are independent for each i, this is the sum of products of independent estimators. This means that covariances across surveys are zero, but can be non zero within surveys. Letting $E_i = P_i\bullet PC_i\bullet CP\bullet CU_i$

$E = \Sigma E_i$ where the sum is over all cells used to define the universe for E. Then

$$\text{Var (E)} = \sum_i \text{Var}(E_i) + \sum_{i\neq j}\sum Cov(E_i,E_j).$$

$$E[E_i] = E\ [P_i]\bullet E[PC_i]\bullet E[CP]\bullet E[CU_i].$$

$$Var(E_i) = E[E_i^2] - E[E_i]^2$$

$$= E\ [P_i^2]\bullet\ E[PC_i^2]\bullet E\ [CP^2]\bullet E[CU_i^2]$$

$$- E[E_i]^2$$

$$= [E[P_i]^2+Var(P_i)][E[(PC_i]^2+Var(PC_i)]$$

$$[E[CP^2] + Var(CP)]\ [E[CU_i^2] + Var(CU_i)]$$

$$- (E[P_i]\bullet E[PC_i]\bullet E[CP]\bullet E[CU_i])^2$$

A similar formula can be developed for Cov (E_i, E_j).

For two independent variables x and y an approximation is [Kish]

$$\text{Var}(xy) \approx \text{E}[x]^2\ \text{Var}(y) + \text{E}[y]^2\ \text{Var}(x).$$

For 4 terms the expansion is simple. Given w,x,y and z independent

$$Var(wxyz) \approx (E[x]\bullet E[y]\bullet E[z])^2\ Var(w)\ +$$

$$+\ (E[w]\bullet E[x]\bullet E[z])^2\bullet Var(y)$$

$$+\ (E[w]\bullet E[y]\bullet E[z])^2\bullet Var(x)$$

$$+\ (E[w]\bullet E[x]\bullet E[y])^2\ Var(z).$$

Similar approximations exist for covariances. In practice the expected values are approximated with their sample values. To estimate these terms we simply evaluate the conditional variances of each of the terms given the others. For instance, $x^{2} \cdot y^{2}\, z^{2}\ Var[w]$ the estimate for the first term above is $Var_{w}\ [wxyz/xyz]$. Thus we may approximate $Var\ [wxyz]$ as

$$Var[wxyz] = Var_{w}\ [wxyz/xyz]$$

$$+\ Var_{y}[wxyz/wxz]$$

$$+\ Var_{x}\ [wxyz/wyz] + Var_{z}[wxyz/wxy]$$

For a sum of two products $x_1y_1 + x_2y_2$ of independent variables

$$Var[x_1y_1+x_2y_2] = Var[x_1y_1] + Var[x_2y_2] + 2\ \mathrm{Cov}\ (x_1\ y_1,\ x_2\ y_2)$$

Using these formulas for approximate variances and covariances, we can show that an approximation of $Var(E) = Var(\Sigma E_i)$ is

$$\sum_{i} [VarE/PC_i\text{'}s,CP,CU_i\text{'}s)$$

$$+\ Var(E/P_i\text{'}s,PC_i\text{'}s,CU_i\text{'}s)$$

$$+Var\ \ (E/P_i\text{'}s,CP,CU\text{'}s)$$

$$+\ Var(E/P_i\text{'}s,PC_i\text{'}s,CP).$$

which is the sum of the conditional variances of E over each of the surveys given the others are fixed.

Empirical Results

In order to examine the variances of sums over many cells and to examine relationships of variances for different models we calculated 7 models for each of 3 different types of expenditures. The models were created by breaking the population into finer and finer sets of demographic cells. The variables used to define these cells were: age, 5 levels, sex, 2 levels, poverty status, 3 levels, insurance coverage, race, 2 levels, and household size, 2 levels.

By letting the first model contain only 1 cell and progressively adding variables we created models with 1, 5, 10, 30, 90, 180 and 360 cells. This essentially created a series of model estimates each which assumed another variable related to health care expenditures.

For three types of expenditures and the total of the three expenditures: 1. outpatient physician visits, 2. emergency room physician visits, and 3. all other physician visits we created estimates, if the model allowed, (A model which did not use a given variable to define the model was not used to create estimates for subpopulations defined by that variable. For instance the 5 cell model which only was defined by age was not used to make estimates of totals by sex.) for the total population and 17 subpopulations: total population (1), total by sex (2), total by age and sex (10), and total by age (5).

This allowed us to empirically compare estimates and their relative errors for the different models. Estimates of variances were calculated using the computer package SURREGR [Holt, 1977] for the NMCES and NHIS surveys. For the CPS variances we used generalized variance formulas. For the CPI we used results for all health care [Leaver, et. al, 1991] and prorated variances to subindexes according to their relative importances. In general the CPI and CPS variances were smaller and contributed little to overall variances.

Tables A, B and C show a set of representative results using the 7 models to derive health expenditure estimates for the entire population, for the following health care services, other physician visits, outpatient visits and emergency room visits. From these results one can observe three patterns.

1. The standard errors, as one might expect, are greater for model estimates than for an individual survey. For instance, for other doctor visits for the entire population, the NMCES relative error is about 1.8% versus the 2.85% for the model error.

2. The large changes in standard errors for different models seem to correlate positively with large changes in the estimates. For

instance, for outpatient visits for the entire population, the largest movement in standard errors is between the 1 and 5 cell model which is when the largest movement in the estimate occurs. This pattern occurs to a lessor extent for many of the other estimators.

3. When there is little change in the estimates, such as, with other physician visits, there can still be a pattern of change in the standard errors, but the direction fluctuates and the relative sizes of the changes are less than those where there are large changes in the estimates.

While some of these differences can be caused simply by the variances in the estimates themselves, one may analyze the theoretical variances to show that one should expect such patterns. We will show these results in the following section.

Standard Error Analysis

The reason the standard errors of the estimates are larger than those of the individual survey is easy to see if one examines the results for a model with a single cell. For a single cell estimate we have seen that

$$Var(E_i) \approx E[P_i \bullet PC_i \bullet CP]^2 \bullet Var(CU_i)$$

$$+ E[P_i \bullet PC_i \bullet CU_i]^2 \bullet Var(CP)$$

$$+ E[P_i \bullet CP \bullet CU_i]^2 \bullet Var(PC_i)$$

$$+ E[PC_i \bullet CP \bullet CU_i] \bullet Var(P_i).$$

Dividing by $E[P_i \bullet PC_i \bullet CP \bullet CU_i]^2$ *gives*

$$(RE(E_i))^2 = (RE(P_i))^2 + (RE(PC_i))^2 + (RE(CP))^2$$

$$+ (RE(CU_i))^2$$

Thus, the square of the relative standard error for the single cell model is equal to the sum squares of the relative standard errors of each of the surveys used. This relationship combined with other relationships between standard errors for models with differing numbers of cells may be useful in helping make rough approximations of relative errors without having to perform the actual extensive variance calculations involved.

We can examine the relationships between models by using the structure of generalized variance formulas published with many surveys [Schoenborn and Marano, 1988] and the linearized approximation used to approximate variances. Consider a single cell model and an associated multicell model. For simplicity we will assume only results from two surveys are multiplied together. One survey provides a population, the other a ratio, such as, a per capita.

The estimate for the single cell model is

$E = a_T \bullet X_T$ where

X_T is the total population $= \Sigma\, x_i$

a_T is the weighted ratio $= \Sigma\, p_i a_i$

where x_i is the population for the ith cell and a_i is the per capita for ith cell and p_i is a weight for the ith cell, $\Sigma\, p_i = 1$.

The estimate for the multicell model is

$E_m = \Sigma\, a_i x_i$.

From generalized variances we may show the following

$$Var\ (x_i) = bx_i^2 + cx_i$$

$$Cov(x_i,\ x_j) = bx_i\ x_j$$

$$Var(a_i) = a_i^2\ (d + e/s_i)$$

$$Cov\ (a_i,a_j) = a_j\ a_j\ (d)$$

where b, c, d, e, f are constants and s_i is a size variable related to the cell. For instance, for a per capita from the 1977 NMCES, the size variable would be the 1977 cell population.

Using these results and the formulae derived earlier using conditional variances

Var $(E) = \text{Var}(E/x_T) + \text{Var}(E/a_T)$ and
Var $(E_m) = \text{Var}(E/x_i\text{'s}) + \text{Var}(E/a_i\text{'s})$.

Substituting we obtain

$$\text{Var}(E) \approx a_T^2\ (bx_T^2+cx_T)+x_T^2(a_T^2)(d+e/s_T)$$

$$Var(Em)\ \approx b\ \Sigma\ \Sigma a_i a_j x_i x_j + c\Sigma a_i^2 x_i$$

$$+d\Sigma\Sigma a_i a_j x_i x_j\ +\ e\ \Sigma\ x_i^2 a_i^2/s_i$$

Note $\Sigma\Sigma a_i a_j x_i x_j\ =\ (\Sigma a_i x_i)^2\ =\ E_m^2$

We may examine the differences of Var(E) and Var(E_m) term by term.

$$Var(E)\ -\ Var(E_m)\ =\ (b+d)(E^2\ -\ E_m^2)$$

$$+\ c[x_T[a_T^2\Sigma\ a_i^2\bullet x_i/x_T)]$$

$$+\ e\ [x_T^2 a_T^2/s_T-\Sigma\ a_i^2\bullet x_i^2/s_i]$$

As one can see, the first term varies directly with changes in the estimate whereas the other two terms all measure variabilities among the values of the a_i's weighted either by the values of s_i, or x_i. Thus the first term shows why large changes in expected values relate to rates of changes; whereas, even with approximately equal expected values, slight variations in the relationships of the variables can cause different patterns in variances.

One can demonstrate the potential effects of the latter terms with an example. If one considers a two cell model where $s_i = x_i$ and $x_1 = 1 + \epsilon$, $x_2 = 1-\epsilon$. Then if $a_1 = 1 + \epsilon$, $a_2 = 1- \epsilon$ or if $a_1 = 1-\epsilon$ $a_2 = 1+ \epsilon$, the change in expected values have the same absolute change. For case one $E[E_m] = 2 + 2\epsilon^2$, for the second case $E\ [E_m] = 2 - 2\epsilon^2$. However, the variances can be quite different. (In these generalized variances, the values of c and e are much larger than b and d.) [Waite, 1991, Schoenborn, et. al., 1987]

For the first case, the difference

$$Var(E)-Var(E_m)=(b+d)(-8\epsilon^2-4\epsilon^4)+c(-6\epsilon^2)$$
$$+e(-6\epsilon)^2$$

For the second one

$$Var(E)-Var(E_m)=(b+d)(8\epsilon^2-4\epsilon^4)+c(2\epsilon^2)$$

$$+e(2\epsilon)^2$$

Since c and e are generally greater than b and d for the first example Var $(E_m) >$ Var (E). The opposite is true for the second case. It seems that if expected cell values for the surveys are positively correlated, the variances for the models with more cells will rise, even with equal expected values. Thus, final relationships of variances of models also relate the correlations of the cell estimates for the individual survey estimates.

Conclusions and Recommendations

As we have seen, the relative error for these multicell models have several features:

1. They vary with the expected values, much the way variances for certain totals from standard surveys vary with the size of their expected values.
2. They are larger than the variances for a similar estimate from a single survey. For a single cell model the variances can be estimated using relative errors from each survey
3. For models with large numbers of cells, variances of these estimates are very expensive to calculate because of the need for variances, covariances and estimates from each individual

survey.

However, there may be alternative estimates. If one had generalized variance formulas for each survey one could try to construct a generalized variance for this estimator or to approximate the RSE. Another useful approximation might be to use the single cell approximation of RSE.

Although it would be best to calculate actual errors, if costs prevent this being done, efforts should be made to provide policy makers with approximations so that potential accuracy of these estimates are known.

(references upon request)

Table A: Other Physician Visits

Group	Cells in Estimate	No. of Cells in Model	Estimate/10^{10}	Std. Error/10^9	RSE X 100	NMCES x 100 RSE
All	1	1	4.233	1.206	2.85	1.79
	5	5	4.272	1.212	2.84	
	10	10	4.274	1.216	2.85	
	30	30	4.274	1.203	2.81	
	90	90	4.276	1.204	2.82	
	180	180	4.247	1.184	2.79	
	360	360	4.275	1.138	2.66	

Table B: Outpatient Physician Visits

Group	Cells in Estimate	No. of Cells in Model	Estimates/10^9	Std. Error./10^9	RSE x 100	NMCES x 100 RSE
All	1	1	6.864	.5945	8.66	6.09
	5	5	7.965	.8125	10.20	
	10	10	7.938	.7659	9.65	
	30	30	8.313	.8541	10.27	
	90	90	8.354	.8777	10.51	
	180	180	8.436	.9410	11.15	
	360	360	9.160	1.1201	12.21	

Table C: Emergency Room Physician Visits

Group	Cell in Estimate	No. of Cells in model	Estimates/10^9	Std. Error/10^9	RSE x 100	NMCES RSE x 100
All	1	1	5.048	.3826	7.58	3.58
	5	5	5.317	.4065	7.64	
	10	10	5.378	.4138	7.69	
	30	30	5.513	.4527	8.21	
	90	90	5.445	.4462	8.20	
	180	180	5.731	.4913	8.57	
	360	360	5.756	.5095	8.85	