NATIONAL HEALTH INTERVIEW SURVEY FAMILY WEIGHTING RESEARCH 2000

Karen E. Davis, National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782

Key Words: NHIS, Family weight

Introduction

The National Health Interview Survey (NHIS) is one of the major data collection programs of the National Center for Health Statistics (NCHS). Through the NHIS, information concerning the health of the U.S. civilian noninstitutionalized population is collected in household interviews throughout the U.S. Beginning with the 1997 NHIS, a family file with family weights was produced due to the questionnaire restructuring. For example, distributional estimates for families, such as the percent of families where at least one member has at least one hospital stay, can now be estimated. This paper describes research undertaken to compare the current NHIS family weighting procedure to four alternative family weighting methods.

Regional estimates obtained by using five different family weighting methods were compared to independent population controls prepared by the U.S. Bureau of the Census. The five methods were: arithmetic mean of person weight (AMP), geometric mean of person weight (GMP), generalized least squares (GLS), raking ratio estimation (RAKE), and the current (1997) procedure which is a variant of the "principal person method " (Alexander, 1987).

We begin with a general description of the NHIS weighting process. This is followed by a description of the alternative methodologies that were employed in the research. We end by presenting results with some discussion of future work.

Background

The 1997 NHIS collected data on 103,477 persons within 40,623 families living in households. The 1998 NHIS collected data on 98,785 persons within 38,773 families living in households. The 1997 and 1998 Family Files are considered household-level files for single family households. Note that most households contain only one family. However, for multiple family households, each record in the Family File represents a unique family. Since the NHIS does not sample families as a stage of sampling, family weights must be derived by an indirect process. The weights used to create the current family weight originate from the person weights in the Person File. The person weight is a product of up to four weighting factors:

- inverse of the probability of selection
- household nonresponse adjustment
- first-stage ratio adjustment (by race, ethnicity, region, and metropolitan area status)
- second-stage ratio adjustment (post-stratification to 88 age, race, ethnicity, and gender Census control totals)

For estimates at the household-level, the sampling weights in the NHIS Household File are created from the person weights. The household weight is defined to be the person weight up to and including the nonresponse adjustment.

The 88 post-strata are determined by the following 44 groups for males and females:

Hispanic	**Non-Hispanic Black**	**Other**
Under 1 year	Under 1 year	Under 1 Year
1-4 years	1-4 years	1-4 years
5-9 years	5-9 years	5-9 years
10-14 years	10-14 years	10-14 years
15-17 years	15-17 years	15-17 years
18-19 years	18-19 years	18-19 years
20-24 years	20-24 years	20-24 years
25-29 years	25-29 years	25-29 years
30-34 years	30-34 years	30-34 years
35-44 years	35-44 years	35-44 years
45-49 years	45-49 years	45-49 years
50-54 years	50-54 years	50-54 years
55-64 years	55-64 years	55-64 years
65+ years	65-74 years	65-74 years
—	75+ years	75+ years

The current NHIS family weighting procedure is a variation of the principal person method. In the existing procedure, the family weight for a particular family record is the final weight of the person in the family with the smallest post-stratification factor.

For person weights, post-stratification provides a way to reduce both bias and variance of estimates since it assures that the NHIS estimates agree with independently determined Census population controls. These controls include an adjustment for net undercoverage in the 1990 Census, and are used for the Current Population Survey (CPS). Unfortunately, for family weights, no suitable independent estimates of the number of families exist from an external source, such as the U.S. Bureau of the Census, to perform post-stratification.

The GLS weighting method obtains weights by minimizing a GLS objective function subject to the constraint that person estimates obtained by using family weights agree with person control totals (Ikeda, 1993). One problem with this method that may occur is that the final weights can be negative (Peitzmeier et al., 1988).

The AMP and GMP are weighting methods that provide alternatives to using a principal person weight (Navarro et al., 1991). Both of these methods use the average weight of the person weights of all persons considered to make up a family. This average weight can then be used for tabulating family data. The AMP weight is an unbiased estimate of the total. Both the AMP and GMP methods apparently reduce bias without increasing the variance of the estimates (Navarro et al., 1991).

The RAKE method is an iterative procedure which can be used to assure consistency between complete count and sample data. This method uses a contingency table structure and assumes that two or more marginal population totals of the contingency table are known, and that the interior of the table can be estimated from the sample. After a specified number of iterations are completed or when the sums are simultaneously satisfied to the closeness desired, a raking factor is determined. This factor can be applied to the initial weights to produce a final family weight. Raking estimates, although not maximum likelihood estimates under certain conditions, are consistent and best asymptotically normal (Mulrow et al., 1993).

Methods

The 1997 and 1998 NHIS Person Files and Family Files were used in this research. These files were sorted by families within households.

The first alternative method was the AMP. For this method, the weight is calculated as the mean weight of the person weights of all persons considered to make up the family. It is the expected value of the weight assuming a member of the family is selected with equal probability to represent the family. It is defined as:

$$(1)\ (\mathrm{AMP})_i = \sum_{j=1}^{a_i} W_{ij} \ / \ a_i$$

W_{ij} is the final annual weight (WTFA) of the j th person in the i th family, a_i is the number of family members in the family.

The use of the arithmetic mean weight produces an unbiased estimate of the total.

The second method was the GMP which is based on the geometric mean of person weights. Asymptotically, the geometric mean is close to the median for moderate to large numbers, and is in some sense a typical value (Hines, 1983). The geometric mean of person weights can therefore be considered a median value of the weight to represent the family. Also, a known result is that the geometric mean is always less than or equal to the arithmetic mean for positive variables. It is defined as:

$$(2)\ (\mathrm{GMP})_i = \prod_{j=1}^{a_i} W_{ij}^{\,1/a_i}$$

W_{ij} is the final annual weight (WTFA) of the j th person in the i th family, a_i is the number of family members in the family.

The geometric mean appears to be a bias reducing method, but which apparently does not increase the variance of the estimates (Hines, 1983).

The third method was the GLS weighting method. The goal of the GLS is to provide consistency between family estimates and estimates for the full population, by producing family weights that agree with independent demographic person controls, and are as close as possible to the initial weights assigned to the persons in families. The GLS minimizes a least squares objective function with respect to a set of initial person weights subject to the constraint that estimates obtained by using family weights agree with the Census population totals. The vector of family weights is defined:

$$(3)\ \overline{GLS} = \overline{S} + MA(A'MA)^{-1}(\overline{N} - A'\overline{S})$$

The NHIS Person File contains the interim annual weight (WTIA) which does not include the post-stratification adjustment. $\overline{S}$ is

the vector of interim annual weights. $\overline{N}$ is the vector of 88 post-stratification cells with the control counts. The sample households comprise matrix A= (a_{ik}) where a_{ik} is the number of persons in the i^{th} family in the k^{th} post-stratification cell. Lastly, M is a diagonal matrix with values $S_1/a_{1.}$,…, $S_i/a_{i.}$.

The fourth method was the RAKE method. The goal of the RAKE, like the GLS, is to provide consistency between sample estimates and estimates for the full population, by producing family weights that agree with independent demographic person controls, and are as close as possible to the initial weights assigned to the sample persons. The RAKE method is constrained so that estimates obtained by raking usually assume that two or more marginal population totals, say $N_{i.}$ and $N_{.j}$, are known. The raking algorithm begins by setting $N_{ij}=(N/n)n_{ij}$, and then proceeds by proportionately scaling the N_{ij} such that the relations

$$(4)\ \sum_j N_{ij} = N_{i.}$$

and

$$(5)\ \sum_i N_{ij} = N_{.j}$$

are satisfied in turn. Each step in the algorithm begins with the results of the previous step, with the N_{ij} continuing to change. At the end of this process, the raking weight is defined as:

$$(6)\ (RW)_{ij} = F_{ij}(IW)_{ij}$$

where F_{ij} is the raking factor needed to reweight the sample so that it agrees with the known population totals, and $(IW)_{ij}$ is the initial sample weight.

For the RAKE method, NHIS data were classified into 2 x 44 = 88 cells based on gender, age, race, and ethnicity, corresponding to the post-stratification cells for raking estimation. An initial family weight was calculated as the minimum interim annual weight (WTIA) for each of these 88 cells. An observation was then created representing families with the same composition (e.g. Hispanic females age 45 to 49 years), but different family sizes, for each of the 88 cells.

The fifth method was the current NHIS weighting method which is a variant of the principal person method. This method was implemented in the 1997 sample after empirical research by a colleague Christopher Moriarity in the Office of Research and Methodology, NCHS, that indicated that some form of ratio adjustment was advisable, but methods such as the principal person method appeared to adjust too much. This method incorporates the complex NHIS sample design because the final annual weight of the family member with the smallest post-stratification factor is selected as the family weight:

$$(7)\ W_f = W_I \bullet W_{nr} \bullet W_{r1} \bullet \min(W_{r2})$$

W_I is the inverse of the probability of selection, W_{nr} is the household nonresponse adjustment, W_{r1} is the first-stage ratio adjustment, and W_{r2} is the second-stage ratio adjustment (Botman et al., 2000).

Unfortunately, for family weights, no suitable independent estimates of the number of families exist from an external source, such as the U.S. Bureau of the Census, to perform post-stratification. Therefore, for evaluation, each family weighting method was used to calculate regional estimates for total households, since the available Census data was for households. The ratio of the Estimate/Census was then calculated under each alternative to determine which of the methods were closest to the independent controls prepared by the U.S. Bureau of the Census.

Results

Tables 1 and 2 below summarize the results by comparing the Census total households with the current family weighting method, the AMP method, the GMP method, and the RAKE method by Census region. Table 3 compares the 1997 and 1998 NHIS weighted household estimates to estimates from the current method.

Using 1997 and 1998 NHIS data, the AMP, GMP, and RAKE methods generally overestimate the number of total households, while the current family weighting method appears to give estimates closest to Census results. The only exception is in the West region, where the RAKE method provides the best estimate of total Census households, and the current method gives the largest underestimate. Note, however, that since post-stratification in the NHIS is implemented at the national level, regional alignment should not necessarily be expected to occur.

The GLS method was resource intensive and required an inordinate amount of

computer memory for matrix calculations. As a result, in order to compute the family weight, NHIS data for 1,000 families were used. The GLS procedure generated 8 negative weights out of these 1,000. This method is therefore not recommended for producing NHIS family weights.

Discussion

In summary, the current family weighting method is recommended for family level analysis of NHIS data. The results for 1997 and 1998 indicate the existing method provides consistent estimates at the regional level for households; presumably, this would also hold for families. In addition, by using the post-stratification factors, the current method incorporates the complex NHIS sample design as a basis for family estimates. This method can also be easily adapted to other complex surveys.

In the future, if the U.S. Bureau of the Census produces suitable independent household or family estimates, research should be undertaken to evaluate their use in the NHIS family weighting process. Further research could also be undertaken to investigate calibration methods that avoid negative weights.

References

1. Alexander, C.H. (1987). "A Class of Methods for Using Person Controls in Household Weighting", *Survey Methodology*, 13:183-198.

2. Botman, S., Moore, T. F., Moriarity, C., Parsons, V. (2000). "Design and Estimation for the National Health Interview Survey, 1995-2004", *Vital Health Statistics, Series 2, Number 130*, 14-19.

3. Hines, W.G.S. (1983). "Geometric Mean", *Encyclopedia of Statistical Science*, 3:397-400

4. Ikeda, M. (1993). "Comparison of Alternative Family Weighting Methods for the National Health Interview Survey", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 770-775

5. Mulrow, J., Oh, H. L. (1993). "Raking Ratio Estimation over Time", *Statistics of Income: Turning Administrative Systems into Information Systems, Internal Revenue Service*, 101-107

6. Navarro, A., Griffin, R. A. (1991). "Family Estimates Empirical Study", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 789-794

7. Peitzmeier, F., Hughes, A. L., Hoy, C. E. (1988). "Alternative Family Weighting Procedures for the Current Population Survey", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 437-442

Table 1. Ratio of 1997 National Health Interview Survey Estimates to Census Total Households for Four Alternative Weighting Methods

	Current Method[1]	AMP[2]	GMP [3]	RAKE[4]	Census Households
U.S.	1.0204	1.0633	1.0617	1.0523	99,883,746
Northeast	1.0473	1.0905	1.0888	1.0698	19,363,610
Midwest	1.0772	1.1183	1.1168	1.0858	23,642,324
South	1.0359	1.0815	1.0796	1.0485	35,448,780
West	0.9075	0.9481	0.9465	1.0056	21,429,032

Table 2. Ratio of 1998 National Health Interview Survey Estimates to Census Total Households for Four Alternative Weighting Methods

	Current Method[1]	AMP[2]	GMP[3]	RAKE[4]	Census Households
U.S.	1.0244	1.0641	1.0625	1.0600	101,041,243
Northeast	1.0387	1.0787	1.0772	1.0680	19,449,612
Midwest	1.0924	1.1306	1.1292	1.0909	23,761,471
South	1.0320	1.0729	1.0711	1.0538	35,984,972
West	0.9250	0.9642	0.9626	1.0295	21,845,188

Table 3. Ratio of 1997 and 1998 National Health Interview Survey Household Estimates to Current Method

	1997 Current Method [1]	1997 Household Estimate	1998 Current Method[1]	1998 Household Estimate
U.S.	1.0204	0.9035	1.0244	0.8913
Northeast	1.0473	0.9310	1.0387	0.9059
Midwest	1.0772	0.9480	1.0924	0.9447
South	1.0359	0.9186	1.0320	0.8978
West	0.9075	0.8048	0.9250	0.8093

Note: Post-stratification in the NHIS is implemented at the national level, so regional agreement should not be expected to occur.

[1] Current method calculates the family weight using the final annual weight of the family member with the smallest post-stratification factor.
[2] The AMP method calculates the family weight as the arithmetic mean of all the person weights in the family.
[3] The GMP method calculates the family weight as the geometric mean of all the person weights in the family.
[4] The RAKE method calculates the family weight using raking ratio estimation.