GENERALIZED VARIANCE FUNCTIONS FOR THE 1994 NATIONAL EMPLOYER HEALTH INSURANCE SURVEY

Christopher L. Moriarity, Sarah W. Gousen, David W. Chapman
Christopher L. Moriarity, National Center for Health Statistics,
6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Variance estimation

1. Introduction

The National Employer Health Insurance Survey (NEHIS) was conducted in 1994 by Westat, Inc., under contract to the National Center for Health Statistics (NCHS). The purpose of the NEHIS was to collect information on the health care insurance that U.S. businesses and governments provide for their employees. The survey collected information from employers on the names and types of health insurance plans (if any) offered to their employees, enrollments in these plans, the characteristics of the plans, the money paid for claims in the preceding plan year, and other related data.

The target sample size for the 1994 NEHIS was about 37,000 interviews for private establishments (i.e., specific business locations) and about 3,000 interviews for government agencies, for a total of about 40,000 interviews. The sample design was a stratified random sample of establishments. Strata were defined by state and size class in terms of the number of employees in the establishment. In the private sector, the number of employees in the "firm" containing the establishment also was used as a stratifier. For the public sector, type of government was included in the stratification process. In general, establishments in larger size classes were sampled at higher rates. An overview of the sample design is provided by Marker, et al. (1994). Additional details of the sample design, including the sources of the sampling frames, are provided by Westat (1994, 1996). An overview of the weighting and estimation procedures is given in Wallace, et al. (1995).

NCHS is preparing publications from the NEHIS data. The NEHIS publications will not include standard errors for most of the estimates given. Instead, the NEHIS publications will include generalized variance functions (GVFs). These functions allow users to compute approximate variances for estimates.

This paper summarizes our research on various alternatives for providing GVFs for the first NEHIS publication. We describe the regression models chosen and the criteria used in the selection process. We compare our methodology with procedures used to produce GVFs for other surveys at NCHS, such as the National Health Interview Survey.

2. GVFs for NEHIS totals and percents

Our general approach for finding GVFs for the first NEHIS publication was similar to the procedure outlined in Wolter (1985, p. 205). First, we obtained direct estimates of sampling errors, then we did some data analysis and formed subgroups. Next, we used the direct estimates of sampling errors as the dependent variable in models, where we fit a different model for each subgroup. We used SUDAAN (Shah, et al., 1996) to obtain design-based direct estimates of sampling errors, and we used SAS procedures and SAS data step programming to do modelling. Most of our modelling used ordinary least squares, although we also examined the result of fitting some models using weighted least squares or iteratively reweighted least squares. The ordinary least squares modelling was an iterative process, with up to six iterations. Observations with a studentized residual greater in absolute value than 3.5 were discarded as "outliers", and the model fitting was repeated in the next iteration.

The NEHIS publication does not contain the number of sample cases that contributed to a given estimate. For this reason, we did not use the sample size as a regressor for GVF models.

The NEHIS publication has tables containing estimates of percents and totals, along with a few estimates of means and a few "standardized" estimates. We decided to limit our model fitting to percents and totals, using models for percents that differed from those used for totals.

In addition, NEHIS estimates can be classified into two broad groups: "employee-related" and "establishment-related". An example of an employee-related estimate is the percentage of employees who obtain health insurance through their employer. An example of an establishment-related estimate is the percentage of business establishments that offer health insurance to their employees. Although we used the same general model equation for all totals, we fit separate model parameters for employee-related estimates versus establishment-related

estimates. We followed a similar strategy for employee-related percents versus establishment-related percents. We decided to group estimates by employee-related versus establishment-related after doing data analyses such as scatterplots, and fitting overall models versus fitting separate models by employee-related estimates versus establishment-related estimates.

We also did some exploratory work where we gave consideration to other groupings of estimates in addition to employee-related versus establishment-related.

Historically, GVFs have been fit using a limited number of direct estimates of sampling error. We decided to exploit the rapid advances in computing power that have occurred in recent years by computing a complete set of direct estimates of sampling error for all estimates of percents and totals that are included in the NEHIS publication. Hence, the GVFs we created do not contain error that is due to subsampling of direct estimates of sampling error for developing the GVFs.

As indicated in the introduction, the sampling units in the 1994 NEHIS were business establishments in the private sector, and governments in the public sector. The distribution of employment in the businesses and governments is strongly skewed to the right; that is, most businesses and governments have a small number of employees, while a relatively small proportion of businesses and governments have large numbers of employees. Hence, the distribution of the number of employees in businesses and governments is quite different from the distribution of persons in households. Therefore, it would not be surprising to find that a generalized variance function model that performs well for population surveys might not perform too well for an establishment survey.

2.1 GVFs for totals

The Current Population Survey, and NCHS surveys such as the National Health Interview Survey, employ the GVF model

$$\text{relvar}(x) = a + b/x \qquad (1)$$

to model the relative variance (i.e., the variance divided by the square of the estimate) of a total "x" as a linear function of 1/x. This model has some theoretical justification; see Wolter (1985, pp. 203-204). Valliant (1987) also provides some theoretical justification for this model.

Our starting point for modeling GVFs for totals was Equation (1). However, the scatterplot of relvar(x) versus 1/x did not show a linear pattern. We were not satisfied with the performance of Equation (1), as measured by the R^2 value and a scatterplot of predicted standard errors versus direct estimates, so we tried other models. The models we tried included other models given in Wolter (1985, p. 203, Equations 5.2.2 - 5.2.5), a few others of similar form, and a model mentioned in Kalton (1977, p. 506). None of the models gave what we considered to be a satisfactory fit to the data. A serious problem with one of the models (Wolter, 1985, Equation 5.2.4) is that it gave negative estimates for some relative variances for observed values of x. We decided to reject any model that displayed this behavior.

We also tried a loglinear model that does not appear in Wolter (1985):

$$\ln\{\text{var}(x)\} = a + b \ln(x), \qquad (2)$$

where "ln" denotes the natural logarithm. For both employee-related and establishment-related totals, the scatterplot of ln{var(x)} versus ln(x) showed a pattern that indicated a linear model would be a reasonable fit. For both groups of totals, the scatterplot of predicted standard errors versus direct estimates generally followed the y = x line, although there was a "hook" in the scatterplot where the largest values of the predicted standard error corresponded to decreasing values for the direct estimate. The "hook" was quite obvious for establishment-related totals. Although we believed Equation (2) gave a better fit to the data than the models we had tried previously, using the R^2 criterion, we continued to search for other models that might provide additional improvement.

We tried a variation of Equation (1), viz., we multiplied both sides by x^2 to obtain

$$\text{var}(x) = ax^2 + bx \qquad (3)$$

We decided to work with this model after examining a scatterplot of var(x) versus x for establishment-related totals (Diagram 1), which suggested that a quadratic model might be useful. We fit this model both with and without an intercept term. The model with an intercept term turned out to be unacceptable because it gave some negative predicted variances for observed values of x. The model without the intercept term performed well in some ways, eliminating the "hook" in the scatterplot of predicted standard errors versus direct estimates that we had seen when using Equation

(2) as our model. However, there was a tendency for the model to predict considerably larger variances than the direct estimates at the "low end" of the direct estimates.

Although the models given by Equations (2) and (3) appeared to provide improved performance over Equation (1), our research continued in the hope of finding a parsimonious model that provided a good fit to the observed data over the entire data range. We gave consideration to a model that is a generalization of Equation (3):

$$\{var(x)\}^r = ax^2 + bx \qquad (4)$$

We became interested in this model after we examined the results of using the values r = 1 (i.e., Equation (3)), and r = .5 (i.e., se(x) = $ax^2 + bx$). As indicated above, the model with r = 1 had a tendency to overestimate variances at the "low end". We found that the model with r = .5 did not have a tendency to overestimate variances at the "low end", but there was a tendency to underestimate variances in the "middle". We experimented with using a range of r values between .5 and 1; the best results occurred near r = .75, so we decided to use r = .75. This was the model that we used to fit GVFs for totals.

The results, for establishment-related totals, of fitting the model we chose are summarized in Diagrams 2 and 3. Diagram 2 shows the residual plot for the final iteration in the model-fitting process, and Diagram 3 shows the scatterplot of predicted standard errors versus direct estimates from SUDAAN. Diagram 3 includes "outliers" rejected during the iterative model-fitting process.

We made one modification to the GVF model that we fit for establishment-related totals using Equation (4) with r = .75. We noted that the predicted value for $\{var(x)\}^r$ for the U.S. total number of establishments became negative after several iterations of the model-fitting process. We considered several alternatives, including fewer or no iterations. We also considered forcing the U.S. total to remain in the model-fitting process, even though the variance estimate for the U.S. total typically was rejected as an "outlier" in many of the equations we explored for fitting GVFs for totals. We decided that a better overall fit to the data was achieved with the iterative process we were using, and we found that forced inclusion of the U.S. total throughout the iterative process did not prevent the predicted value from becoming negative after several iterations. To remedy the situation, we defined the predicted standard error to be 20,000 for all establishment-related estimates greater than 4.7 million. (Note that the U.S. total estimate from NEHIS of business establishments is approximately 6.3 million, and the direct estimate of the standard error is approximately 8,000.)

2.2 GVFs for percents

NCHS surveys such as the National Health Interview Survey employ the GVF model

$$var(p) = bp(100 - p)/y \qquad (5)$$

to model the variance of a estimated percent "p". This model is generated from an application of the model given by Equation (1) for modeling the variance of an estimated total. This model is discussed in Wolter (1985, p. 204). This model assumes that the denominator "y" of the percent "p" is available, or can be derived, for data users to compute GVFs.

Since we are not using models based on Equation (1) to model totals, we decided to use a slightly different GVF model for percents:

$$\ln\{se(p)/sqrt(p(100 - p))\} = a + b \ln(y), \qquad (6)$$

where "sqrt" denotes the square root. Note that this model, like Equation (5), assumes that the denominator of the percent p is available, or can be derived, for data users to compute GVFs. The NEHIS publication does contain this information.

We were satisfied with the performance of this model, as measured by the R^2 value and a scatterplot of predicted standard errors versus direct estimates. Diagram 4 shows the scatterplot for establishment-related percents, including "outliers" rejected during the iterative model-fitting process.

By taking exponentials of both sides of Equation (6), followed by some algebra, Equation (6) can be written as

$$var(p) = e^{2a} * p(100 - p)/y^{-2b}, \qquad (7)$$

which shows the similarity to Equation (5). "b" in Equation 5 is positive, as is e^{2a} in Equation (7). Typical fitted values for the exponent of y, "-2b", in Equation 7 were in the 0.6 to 0.7 range, which compares to an exponent of 1 for y in Equation 5. An additional similarity between Equations (5) and (7) that is intuitively appealing is that both are symmetric in p

and (100 - p), implying that var(p) = var(100 - p).

Prior to focussing on Equation (6), we did some exploratory model-fitting using models that were similar to Equation (6), but more complicated. These models were of the form

$$\ln\{se(p)\} = a + b \ln(y) + c \ln(p) + d \ln(100 - p) \quad (8)$$

As expected, a better fit is obtained using Equation (8) than Equation (6). However, the improvement in fit was not dramatic, as the estimates of c and d typically were close to .5 (which is equivalent to Equation (6)). Hence, we decided that the slight improvement in fit that we gained was not worth the additional complexity associated with the model in Equation (8). Also, since c was not required to be equal to d during model fitting, the intuitive appeal of symmetry in p and (100 - p) was lost.

Note that the Employee Benefits Survey, a survey conducted by the Bureau of Labor Statistics, uses models similar to Equation (8) to model GVFs for percents. Several recent Employee Benefits Survey publications are included in the References section; these publications contain an appendix that describes the methodology used to produce GVFs for percents.

3. Comparison of models

A variety of dependent variables appeared in the GVF models we explored for totals; e.g., the relative variance, the reciprocal of the relative variance, the variance, and the natural logarithm of the variance. For this reason, use of R^2 could be misleading, as different choices for the dependent variable can lead to different denominators for R^2. For example, it can be shown that the ordinary least squares estimates for Equation (2) and Equation 5.2.5 in Wolter (1985, page 203), which is

$$\ln\{relvar(x)\} = a - b \ln(x) \quad (9)$$

have a definite relationship: the intercepts ("a" values) are equal, and "b" in Equation (2) is equal to "-b+2" in Equation (9), resulting in the same predicted values for var(x). However, the R^2 values need not be equal. For example, for establishment totals, Equation (2) gives an R^2 of approximately 0.8, while Equation (9) gives an R^2 of approximately 0.55. Hence, as indicated in Section 2.1, we thought at one time that we had made progress using Equation (2), when in fact this was not true; we had already obtained the same result using Equation (9).

We decided to employ measures that would allow us to compare models, even if the dependent variable was different. Both measures we used involved taking the difference between each direct standard error estimate and the predicted standard error from the model. One measure ("m_1") was the average of the squared differences, and the other measure ("m_2") was the average of the absolute values of the differences.

Mathematically speaking, m_1 and m_2 are "equivalent metrics"; i.e., $m_1 = 0$ if and only if $m_2 = 0$. However, it need not be true that if m_1 is less for one model than the other, it follows that m_2 is less for that model as well.

Using these measures, Equations (2) and (9) give the same results, as they should. Also note that these measures allow us to assess whether there were larger differences, on average, between direct estimates and predicted values for models that use different transformations of the original dataset. That is, we wanted to be alerted to a situation where a model would fit the transformed data well, but large differences between the (untransformed) predicted values and direct estimates occurred.

Measures other than m_1 and m_2 merit consideration for comparing models; in particular, relative measures (e.g., the average of the relative squared differences). We gave consideration to relative measures akin to m_1 and m_2, but decided that they gave too much influence to the "low end".

One relative measure that we considered briefly, but did not have the resources to investigate, was the average of the absolute values of the difference between each direct standard error estimate and the predicted standard error, divided by the estimate (rather than by the standard error estimate). Such a measure would show the change in the estimated coefficient of variation, an often-used indicator of the reliability of an estimate. We think this measure deserves additional study.

4. Discussion of Methodologies for Fitting GVF Models

As indicated in Section 2, we chose to use ordinary least squares regression for our GVF modelling. Wolter (1985, p. 207) discusses why weighted least squares regression and iteratively reweighted least squares regression might be considered to be preferable strategies. The argument rests on the reasonable assumption that the precision of direct estimates of relative variance gets better as the total "x" increases. Note that NCHS surveys such as the National Health

Interview Survey use iteratively reweighted least squares regression for GVF modelling.

We gave due consideration to using weighted regression strategies. We believe that a weighted regression strategy is the preferred method if GVFs are being fit using a limited set of direct estimates. However, we were working with a complete set of direct estimates of sampling error for all estimates that are included in the NEHIS publication, and we noticed that we had a larger proportion of direct estimates at the "low end". Hence, the use of ordinary least squares seemed appropriate.

5. Summary

For developing GVFs for NEHIS totals, we found Chapter 5 of Wolter to be a helpful guide. An additional useful reference for practitioners is Johnson and King (1987). However, the applied literature in this area appears to be sparse. A search of references found using the Current Index to Statistics did not identify potential GVF models beyond those that appear in Wolter (1985). We did find one additional model in Kalton (1977), as mentioned in Section 2.1.

The first model we investigated for totals was the "standard" GVF model, given in Equation (1). We found that it did not give satisfactory results in terms of providing estimates of variances for 1994 NEHIS totals. Therefore, we investigated the application of several models, including those given in Equations (2), (3), and (4). We ultimately chose to use Equation (4) with $r = .75$. Our selection criteria included scatterplots and use of measures that allowed the dependent variable to vary from model to model. Use of R^2 for model comparison can be misleading if the dependent variable varies from model to model.

For estimating percents, we found a model that provides satisfactory variance predictions, given in Equation (6). This model is a special case of a more general model that we investigated, given in Equation (8). The more general model is similar to models used for developing GVFs for percents for the Bureau of Labor Statistics' Employee Benefits Survey.

The model we chose for GVFs for percents (Equation (6)) has a form similar to the model used for the NHIS (Equation (5)). We illustrated the similarity by deriving Equation (7). Both the NHIS model and the model we have chosen for GVFs for percents have the desirable property of symmetry in p and (100 - p).

Acknowledgement

The authors would like to thank Dale Sanders of the Klemm Analysis Group for his work in computing a complete set of direct estimates of sampling error.

References

Bureau of Labor Statistics (1996). Employee Benefits in State and Local Governments, 1994, Bulletin 2477. U.S. Government Printing Office.

Bureau of Labor Statistics (1996). Employee Benefits in Small Private Establishments, 1994, Bulletin 2475. U.S. Government Printing Office.

Johnson, E.G., and King, B.F. (1987). Generalized Variance Functions for a Complex Sample Survey, Journal of Official Statistics, 3, 235-250.

Kalton, G. (1977). Practical Methods for Estimating Survey Sampling Errors, Bulletin of the International Statistical Institute, 47, 495-514.

Marker, D., Bryant, E., and Moriarity, C. (1994). National Employer Health Insurance Survey Sample Design, ASA 1994 Proceedings of the Section on Survey Research Methods, Volume I, 264-269.

Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1995). SUDAAN User's Manual, Release 7.0. Research Triangle Institute.

Wallace, L., Bryant, E.C., Chapman, D.W., Marker, D.A., and Moriarity, C.L. (1995). Weighting and Estimation Procedures for the 1994 NEHIS, ASA 1995 Proceedings of the Section on Survey Research Methods, Volume I, 192-197.

Westat, Inc. (1994). National Employer Health Insurance Survey Design, a project report submitted to the National Center for Health Statistics, May 5, 1994.

Westat, Inc. (1996). Final Methodology Report, Volume I: Statistical Methodology, a project report submitted to the National Center for Health Statistics, December, 1996.

Valliant, R. (1987). Generalized Variance Functions in Stratified and Two-Stage Sampling, Journal of the American Statistical Association, 82, 499-508.

Wolter, K.M. (1985). Introduction to Variance Estimation, Springer-Verlag, New York.

Diagram 1: Establishment Totals
Variance -vs- Total

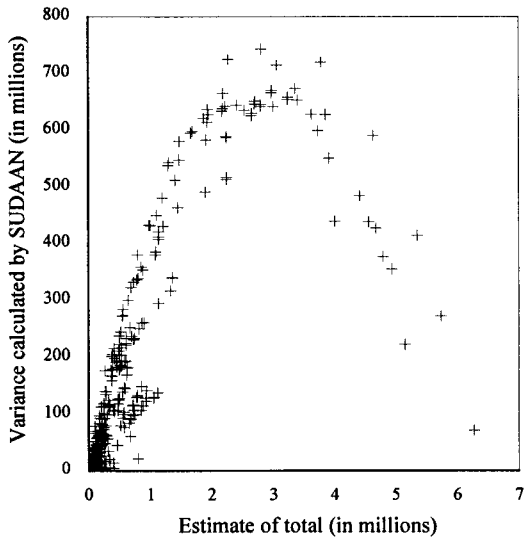

Diagram 2: Establishment Totals
Residual -vs- Total

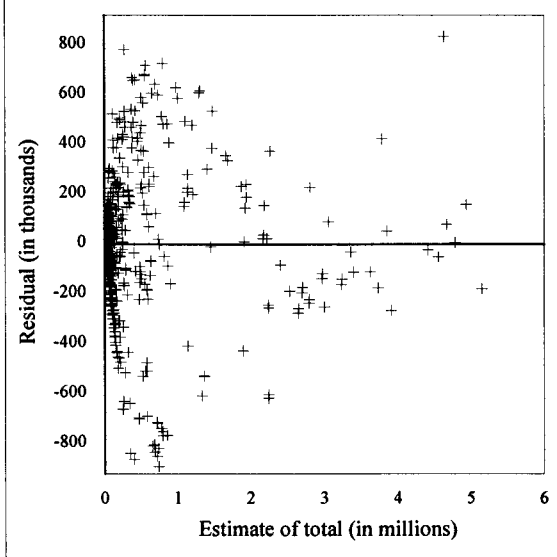

Diagram 3: Establishment Totals
Predicted -vs- Direct estimate of standard error

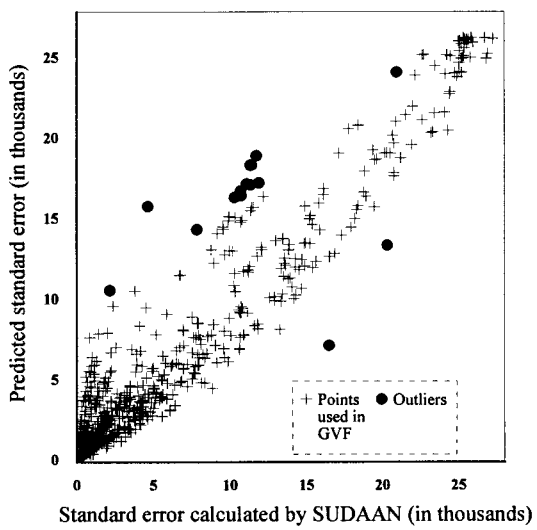

Diagram 4: Establishment Percentages
Predicted -vs- Direct estimate of standard error