AN ASSESSMENT OF CURVE SMOOTHING STRATEGIES WHICH YIELD VARIANCE ESTIMATES FROM COMPLEX SURVEY DATA

Steven B. Cohen, National Center for Health Services Research

1. INTRODUCTION

Complex survey designs, which are generally stratified multi-stage probability designs, require special consideration in the computation of variances. Here, standard methods of variance estimation of sample statistics which assume equal probability sampling are not directly applicable. Often design components include unequal selection probabilities of elements in the population, stratification and several stages of clustering. In addition, the estimation procedures include nonresponse, and poststratification adjustments. In this setting, the assumption of independence is also not valid. Observations made on the sample units are not independent due to the correlation induced by cluster sampling and stratification. Stratification usually results in reduced variance, whereas clustering causes larger positive correlations and increased variance.

Generally, standard methods of variance estimation which assume simple random sampling result in an under-estimate of the true variability, when this approach is directly used with data from a complex survey design. The consequences of using variance estimates derived in this manner in the construction of confidence intervals and for statistical inference is an anti-conservative test. This suggests the probability of rejecting the null hypothesis, even when true, is high. This is a costly bias, particularly when the results of the statistical tests will be used to determine policy. However, use of appropriate estimation procedures, which include the methods of Balanced Repeated Replication and the Taylor series linearization,[1] would be prohibitive with respect to computation time and cost if applied to each parameter estimate of interest. Consequently a curve smoothing technique was developed at the Bureau of the Census and the National Center for Health Statistics [2] which depends on appropriate variance estimates under consideration. One could then determine variance estimates for all related statistics by applying the final prediction equation.

We considered the reliability of the NCHS curve smoothing strategy to data from the National Medical Care Expenditure Survey (NMCES) and whether any improvement could be achieved in precision by application of an alternative strategy which considers weighted least squares. The National Medical Care Expenditure Survey was established to provide a detailed assessment of the utilization, costs, and sources of payment associated with medical care in the United States. The data will meet the needs of government agencies, legislative bodies, and health professionals for more comprehensive national data required for the analysis and formulation of national health policies. The survey was designed to provide data for a major research effort in the Division of Intramural Research of the National Center for Health Services Research (NCHSR), and cosponsored with the National Center for Health Statistics (NCHS).

The area sampling design for NMCES can be characterized as a stratified three-stage area probability design from two independently drawn national area samples.[3] Except for difficulties associated with survey nonresponse and other nonsampling errors, statistically unbiased national and domain estimates can be produced. The essential ingredient of this design is that each sample observation has a known, nonzero selection probability. National general purpose area samples of the Research Triangle Institute (RTI) and the National Opinion Research Center (NORC) were used in NMCES. The structures of both national samples are similar and thereby compatible.

The first stage in both designs consists of primary sampling units (PSU's) which are counties, parts of counties, or groups of contiguous counties. The second stage consists of secondary sampling units (SSU's) which are Census enumeration districts (ED's) or block groups (BG's). Smaller area segments constituted the third stage in both designs from each of which a sub-sample of households was randomly selected in the final stage of sampling. Combined stage-specific sample sizes over the two designs were 135 PSU's (covering 108 separate localities), 1,290 SSU's and 1,290 segments. Sampling specifications required the selection of approximately 13,500 households.

1.2 Variance Curves

Variance estimates were not computed for each statistic considered in NMCES, due to the constraints of computation time and cost. Another consideration was that inclusion of all relevant variance estimates in NMCES data reports would yield rather cumbersome documents. Consequently, we considered the NCHS curve smoothing technique which depends on appropriate variance estimates for only a representative subset of all parameter estimates under consideration. One can then determine variance estimates for all related sample statistics by using the final prediction equation. The subset of statistics that were included in this curve fitting procedure were defined by domains whose underlying demographic characteristics insure a wide range of variability in the parameter estimates.

The curve fitting procedure for aggregate statistics considers the empirically determined inverse relationship between the size of an estimate ($\hat{Y}$) and its relative variance. This relationship is expressed as:[3]

$$\text{Rel Var } (\hat{Y}) = \frac{S^2_{\hat{Y}}}{\hat{Y}^2} = \alpha + \beta / \hat{Y} \qquad (1.1)$$

and estimated as

$$\text{Rel Var}(\hat{Y}) = \frac{S^2_{\hat{Y}}}{\hat{Y}^2} \doteq a + b/\hat{Y} \qquad (1.2)$$

where the regression estimates a and b are determined by an iterative procedure. The relative standard error curve is then derived by taking the square root of relative variance curve

$$RSE(\hat{Y}) \doteq \sqrt{a + b/\hat{Y}} \quad . \qquad (1.3)$$

The iterative procedure considered by NCHS produce estimates of a and b that minimize the squared relative residuals of Rel Var (Y).[3] Consider

$$S = \Sigma_i \left[\frac{V^2_{y''_i} - (a + b/y_i)}{V^2_{y_i}} \right]^2 \qquad (1.4)$$

where $V^2_{y''}$ is the observed relative variance and V^2_y the unknown true relative variance. Starting values of a and b are derived by considering the normal equations for S, where V^2_y is approximated by $V^2_{y''}$. Once the values a, and b, are determined, $\hat{V}_y = a + b/y''$ is computed and substituted for V^2_y. This allows the computation of new estimates of a and b. This procedure continues until

$$\left| \frac{a_j - a_{j-1}}{a_j} \right| \leq 1\% \text{ and}$$

$$\left| \frac{b_j - b_{j-1}}{b_j} \right| \leq 1\%$$

Variances of aggregate statistics were also used to derive variance estimates of percentages, where the numerator was a subclass of the denominator. Here, it can be shown that the relative standard error of a percent estimate $\hat{p}$,

where $\hat{p} = \frac{\hat{Y}}{\hat{T}} \cdot 100$,

takes the form:

$$RSE\ (\hat{p}) \doteq \sqrt{\frac{b}{\hat{T}} \cdot \frac{(100-\hat{p})}{\hat{p}}} \quad , \qquad (1.5)$$

where b is the estimated coefficient determined in the curve fitting procedure for aggregate statistics.[4] Consequently, the variability of percent estimates depends on both the respective population base, T, and the percent value.

Variances of ratio estimators are derived by considering the relationship which specifies the relative variance of a ratio is approximately equivalent to the sum of the relative variances of the numerator and denominator. More specifically, consider the ratio estimator $\hat{R}=\hat{X}/\hat{Y}$, where the numerator is not a subclass of the denominator. This relationship takes the form of

$$\text{Rel Var } (\hat{R}) \doteq \text{Rel Var } (\hat{X}) + \text{Rel Var } (\hat{Y}) \qquad (1.6)$$

$$\doteq a_1 + b/\hat{X} + a_2 + c/\hat{Y}$$

$$\doteq a + b/\hat{RY} + c/\hat{Y}$$

and the relative standard error is approximated by

$$RSE\ (\hat{R}) \doteq \sqrt{a + b/\hat{RY} + c/\hat{Y}} \qquad (1.7)$$

Here, the variability of the ratio estimator is inversely related to the size of the respective population base and the ratio estimate.

1.3 Reliability of Variance Curve Smoothing Procedure

In our study, we examined the reliability of the variance curve smoothing procedure relative to NMCES data. In an attempt to represent the diverse set of aggregate statistics that will be generated from the NMCES data base, three distinct classes of statistics were considered: narrow, medium, and wide range statistics. More specifically, the class of narrow range statistics is determined by data which indicate the presence or absence of a population attribute as measured by 1 or 0 at the individual response level. Similarly, medium range statistics consist of measurements which rarely fall outside the range 0 to 5. Wide range statistics are characterized by data more continuous in nature that have much higher upper bounds.

The class of narrow range statistics was represented by NMCES data which distinguished insured and noninsured individuals with varying types of coverage (e.g., private, Medicaid, Medicare, CHAMPUS, CHAMPVA). Data on the number of dental visits and their respective charges served to represent the medium and wide range classes respectively. In all the above cases, the information obtained refers only to the first quarter of 1977 (January 1 - March 31).

To implement the variance curve smoothing strategy, a number of relevant demographic domains were specified (e.g., defined by age, race, sex, region, marital status, education and relevant cross classifications) and population totals for each criterion variable were estimated. These breakdowns of population estimates include those

most relevant for descriptive and analytical purposes. Several representative samples were drawn from this set of estimates for input to the variance curve smoothing procedure. This was accomplished by ordering the respective estimated population totals and selecting observations via a systematic sampling procedure. To adequately represent the tail ends of the respective distributions of estimated population totals (i.e., observations below the 2nd percentile and above the 96th percentile), observations here were included with certainty. The remaining observations were sampled at 10% and 20% rates, respectively, since only a representative subset of the respective parameter estimates and their relative variances should be essential to establishing a reliable prediction equation. The systematic sampling scheme is particularly appealing since it distributes the sample more evenly over the respective population. Its applicability is further enhanced with no periodic variation evident in the ordered totals. Here the estimated variances of the selected aggregate estimates for input to the curve smoothing procedure were generated via the Taylor series method of variance estimation which is appropriate for complex survey data.

To measure the reliability of this variance estimation technique, we considered the average absolute difference, $\bar{A}_1$, and relative average absolute difference $\bar{A}_2$ between observed and predicted relative variances where

$$\bar{A}_1 = \Sigma_{i=1} \frac{\left| V^2_{y''} - \hat{V}^2_{y''} \right|}{n},$$

$$\bar{A}_2 = \Sigma_{i=1} \frac{\left| V^2_{y''} - \hat{V}^2_{y''} \right|}{n \; V^2_{y''}},$$

$V^2_{y''}$ is the observed relative variance

$\hat{V}^2_{y''}$ is the predicted relative variance, and

n is the number of sample observations.

These measures of reliability can be observed in Table I as they relate to NMCES data on the insured population, dental visits and charges for two different 10% and 20% systematic samples. The results presented closely reflect the variability in validity measures observed in the remaining (complement) set of representative samples.

Within each class of statistics, we noted a wide range of variability exhibited by the specified measures of precision across the respective samples. Further examination revealed that several of these indices differed in value at the $\alpha = .05$ level of statistical significance. For example, consider the deviation in the average absolute difference measures ($\bar{A}_1$) for the 10% samples representing the wide range class of statistics. The difference between an average absolute percent deviation (between predicted and observed relative variances) of 2.75% as compared to 1.40% is significant at the $\alpha = .01$ level using a t-test for the difference in means.

Generally, the observed levels of error fall within acceptable bounds (i.e., $\bar{A}_1 \leq .0601$) when considering the savings in computer time and cost one achieves by using the curve smoothing procedure over the standard methods of variance estimation. A typical standard error computer run using the Taylor series approximation (STDERR program in SAS), or the balanced repeated replication technique on NMCES data can easily cost more than $400.00 and use over 500 seconds of C.P.U. time. Several such computer runs may be necessary to obtain all the relevant point estimates of variances for a specific criterion variable which are estimable from a single variance curve. Alternatively, the curve smoothing procedure usually costs less than $10.00 to run and will use only 3-5 seconds of C.P.U. time. Still, the systematic deviation among the fitted relative variance curves as indicated by our measures of validity is cause for concern, particularly so when they occur in the larger samples where greater stability in the prediction equation would be expected. This is most evident when we observe the divergence in estimated regression coefficients across samples which are presented in Table II. In our limited study, no consistent improvement in precision was noted as sample size specifications were increased. Perhaps this is a function of the inclusion of more observations clustered closer to the mean of the distribution of relative variances as sample size specifications increase, thereby obtaining a closer fit in this central interval but losing precision outside this range, especially at the extremes.

Because the empirical relationship between relative variances and aggregate estimates (totals) is linear in parameters, a least squares technique for estimating α and β in

$$V^2_y = \text{Rel Var } (Y) \doteq \alpha + \beta/Y \qquad (1.8)$$

was also appropriate. Here, we considered the linear transformation $Z = 1/Y$ so that $V^2_y \doteq \alpha + \beta Z$. Examination of the residuals from preliminary regression analysis revealed the variance of V^2_y ($\sigma^2_{(V^2_y)}$) varied inversely as the size of the aggregate estimate (Y) increased (or directly as Z increased), so $\sigma^2_{(V^2_y)} = \frac{K}{Y} = KZ$, thus violating the assumption of homoskedasticity. Use of ordinary least squares in this setting would yield unbiased estimates of the regression coefficients but would not be efficient. Consequently, weighted least squares (WLS) was used to produce the minimum variance unbiased estimates of these coefficients. The appropriate weights, w_y, were of form $w_y = \sigma^2/KZ$, where σ^2 is the constant variance of the transformed observations due to the differential

weighting scheme. Here if we re-expressed relationship (1.8) as $\underset{\sim}{V}_y^2 = \underset{\sim}{Z}\ \underset{\sim}{B}$

where $\underset{\sim}{B} = \begin{bmatrix}\alpha \\ \beta\end{bmatrix}, \hat{b} = \begin{bmatrix}\hat{\alpha} \\ \hat{\beta}\end{bmatrix}$ would be estimated by $\hat{b} = (\underset{\sim}{Z}' \underset{\sim}{W} \underset{\sim}{Z})^{-1} \underset{\sim}{Z}' \underset{\sim}{W} \underset{\sim}{V}_y$ and w_y would be the diagonal elements of $\underset{\sim}{W}$.

Table III presents our measures of reliability applied to the weighted least squares estimation procedure. As expected, the average absolute difference ($\bar{A}_1$) was consistently lower than those observed for the iterative procedure since the minimization criterion considered a similar measure. Many of these observed improvements in precision were significant at the $\alpha = .05$ level of significance as determined by application of paired t-tests. In addition, the range of variability across samples for each class of statistic narrowed markedly for the WLS approach. Here, the linear prediction equation was noticeably more stable across samples. This was most evident by examination of the estimated regression coefficients and their standard errors presented in Table IV. Table V presents the percent improvement, I_i, attained in reliability by using the WLS approach over the iterative regression procedure where

$$I_i = 100 \cdot \left(1 - \frac{\bar{A}_i \text{ (WLS)}}{\bar{A}_i \text{ (Iter)}}\right)$$

One surprising observation was the consistent improvement in the relative average absolute difference ($\bar{A}_2$) when using WLS. Since the minimization criterion of the iterative procedure more closely conforms to this index, the a priori expectation was a loss in precision in $\bar{A}_2$ for the WLS strategy. Consequently, the method of weighted least squares appeared to be a more reliable strategy for predicting variances from NMCES data according to our specified criteria.

1.4 Summary

To conclude, the curve smoothing strategies available for approximating variances of domain estimates from complex survey data serve as appealing alternatives to generating point estimates via balanced repeated replication or the Taylor series method. The estimation strategies employed by NCHS are particularly attractive in their systematic applicability to proportions and ratio estimators in addition to aggregate totals. Since these relationships are direct functions of the relative variance approximation for estimated totals, the accuracy of this prediction equation is a major concern. Here, application of a minimization criterion which considers the sum of squared relative residuals will not consistently provide the optimal curve when additional relevant criteria are simultaneously considered. These criteria include minimum absolute residual deviations, sensitivity to outliers, and distributional properties of estimated regression coefficients. Consequently, the best strategy to adopt requires applicaton of those most relevant regression procedures whose minimization rules reflect these specifications. The regression procedure which most consistently satisfies specified measures of reliability should then be selected.

REFERENCES

1. National Center for Health Statistics: Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution. Vital and Health Statistics. PHS Pub. No. 1000 Series 2-No. 14. Public Health Service. Washington, D.C. U.S. Government Printing Office, March 1975.

2. National Center for Health Statistics: Estimation and Sampling Variance in the Health Interview Survey. Vital and Health Statistics. PHS, Pub. No. 1000 - Series 2, No. 38. Public Health Service. Washington, D.C. U.S. Government Printing Office, June 1970.

3. Cohen, S.B., and Kalsbeek, W.D.: Estimation and Sampling Variances in the National Medical Care Expenditure Survey. Forthcoming.

4. National Center for Health Statistics: National Survey of Family Growth, Cycle 2: Sample Designs, Estimation Procedures, and Variance Estimation. Data Evaluation and Methods Research. Series 2, No. 76. DHEW Pub. No. (PHS) U.S. Government Printing Office, January 1978.

Acknowledgment

The author wishes to thank Ronald J. Hirshhorn for his valuable assistance in computer programming and Polly Estrin for her conscientious typing of this manuscript.

The views expressed in this paper are those of the author, and no official endorsement by the National Center for Health Services Research is intended or should be inferred.