GENERALIZED VARIANCE FUNCTIONS FOR THE SCHOOLS AND STAFFING SURVEYS

Sameena Salvucci, Synectics for Management Decisions, Inc., Glenn Galfond, Price Waterhouse, and
Steven Kaufman, National Center for Education Statistics
Sameena Salvucci, Synectics for Mgmt Dec., 3030 Clarendon Blvd, Ste 305, Arlington, VA 22201

KEY WORDS: Variance estimation, Complex survey designs, Relative variance, iteratively reweighted least squares, Repeated surveys

I. Introduction

This paper presents the results of an empirical examination of relative variances of selected statistics estimated from a complex sample survey. This study looked at the data gathered during the 1987-88 Schools and Staffing Survey (SASS) which was a national survey of elementary and secondary schools conducted by the National Center for Education Statistics (NCES). The target populations for the SASS were school administrators (principals and heads), and classroom teachers in public and private elementary/secondary schools. The survey design consisted of two parallel but essentially separate schemes, one for the public schools and one for private (nonpublic) schools. The components of SASS were (1) Survey of Teacher Demand and Shortage (TDS), (2) Survey of Schools (3) Survey of School Administrators, and (4) Teacher Survey. Approximately 13,000 schools and administrators, 65,000 teachers, and 5,600 Local Education Agencies (LEA's) composed the SASS sample.

NCES prepared eight SASS data files corresponding to the four types of surveys of both public and private schools, each of which contains a set of 48 replicate weights. These weights were designed to produce variances using balanced half-sample variance estimation. However, these replicate weights can be utilized only by users who have half-sample replication software available. The purpose of this task is to develop and test a new procedure using generalized variance functions for approximating the sampling error associated with an estimate of interest.

There were a large number of estimates of interest for the SASS. Estimates of proportions, totals and averages at the national level for various subdomains (i.e., region, school level, minority status, school size, community status and combinations of these) were made. Examples include (1) the total number of administrators who earned a bachelors degree, (2) the proportion of Hispanic students (regardless of race) (3) the number of FTE teachers, and (4) the average hours of teaching basic subjects in private schools.

The school sample was a single stage sample stratified by state by school level in public schools, and state by affiliation by school level in private school. Schools were systematically selected using a probability proportionate to size (pps) algorithm.

Within the first stage school sample, a second stage teacher sample was selected stratified by teacher experience level (teachers with three or fewer years of experience were classified into the new teacher stratum, and all other teachers were classified into the experienced teacher stratum). Within a school, teachers were selected systematically with equal probability.

The goal of this effort was to produce generalized variance functions for each of the Schools and Staffing Surveys. The generalized variances are designed for the user who does not have half-sample replication software available, but requires an approximation to the sampling error associated with his/her estimates of interest.

II. Method of Generalizing Variances

A generalized variance function (GVF) is a mathematical model describing the relationship between the variance or relative variance (relvariance) of a survey estimator and its expectation. If the parameters of the model can be estimated from past data or from a small subset of the survey items, then variance estimates can be produced for all survey items by evaluating the model at the survey estimates, rather than by direct computations.

Denote the estimator of a certain attribute of interest as Xhat and let X = E{Xhat} denote its expectation. Then the relvariance can be expressed as follows:

$V^2 = \mathrm{Var(Xhat)}/X^2$

Most of the GVFs to be considered are based on the premise that the relative variance is a decreasing function of the magnitude of the expectation X.

A simple model which exhibits this property is:

$V^2 = A + B/X$, with $B > 0$. (Model 1)

The parameters A and B are unknown and to be estimated. Experience has shown that Model 1 often provides an adequate description of the relationship between V^2 and X. In fact, the Census Bureau has used this model for its Current Population Survey since 1947.

However, in an attempt to achieve an even better fit to the data than is possible with Model 1, the following are alternative forms of relvariance models which may be considered

$V^2 = A + B/X + C/X^2$ (Model 2)

$\log(V^2) = A + B \log(X)$ (Model 3)

$V^2 = (A + BX)^{-1}$ (Model 4)

$V^2 = (A + BX + CX^2)^{-1}$ (Model 5)

where

V^2 = Relative variance

X = Expectation of the selected survey estimate

A,B,C = Unknown parameters to be estimated

Unfortunately, there is very little theoretical justification for any of the models discussed above. There is some limited justification for Model 1 (Wolter (1985).

III. Technical Approach

As a first step, a pilot test was conducted and based on the pilot test conclusions an exploratory analysis procedure was determined. The findings from the exploratory analysis determined which fitted model was to be used as the GVF.

a. Pilot Test

Step 1: Direct estimates of totals for selected student and teacher headcount variables from the School and the Teacher Demand and Shortage surveys at the national level (by sector and community type) were calculated. These estimates were chosen as a provisional group of similar items to be used for model estimation. A direct calculation of the variance of each of the totals using a balanced half-sample replication technique was used to derive the relvariance and the coefficient of variation (CV). Scatter plots of the log of the estimate versus the log of the CV were used to form "final" groups of statistics that followed a common model. These final groups were formed by simply removing from the provisional group those statistics that appeared to follow a different model than the majority of statistics in the group, and added other statistics, originally outside the provisional group, that appeared consonant with the group model.

As noted in Section II, there is no rigorous theoretical justification for any of the models that relate V^2 to X. Because we were unable to be quite specific about any of the models and their attending assumptions, it was not possible to construct, or even to contemplate, optimum estimators of the model parameters. Discussions of optimality would require an exact model and an exact statement of the error structure of the estimator Vhat^2 and Xhat. In the absence of a completely specified model, we attempted to achieve a good empirical fit to the data (Xhat, Vhat^2) as we considered alternative fitting methodologies.

Step 2: Using the calculated estimates and their CV's, un-weighted nonlinear models using SAS NLIN procedure were fit in order to produce least-squares estimates of the parameters of all five of the relvariance models described in section II above for each of the six subdomains groups (made up of combinations of public/private and urban/suburban/rural). The iterative method specified for the NLIN procedure was the modified Gauss-Newton method which regresses the residuals onto the partial derivatives of the model with respect to the parameters until the estimates converge.

Step 3: The results of the NLIN runs were summarized in terms of the RMSE and bias by quartile.

Step 4: An overlay of the scatterplot of the CV's versus the log of the estimate onto the fitted regression curve was plotted for each of the fitted models described in step 2.

Step 5: Finally, the results of steps 3 and step 4 were examined to help determine a viable subset of models to be used for the overall analysis. This determination was made by looking at both how well the data fit the model and how well the shape of the curve was in accord with reality.

Preliminary Results:

Both models 2 and 5 produced inappropriate shapes for the regression curve fit to the data in terms of a danger that extrapolation could lead to a result that was far from in accord with reality. Of the remaining models (1, 3 and 4), model 1 was the worst because the shape of the regression curve often dropped off too fast and leveled off too quickly. The shape of the curve for Model 3 seemed reasonable and appeared to fit fairly well overall, but had a higher RMSE than model 4. Also, model 3 resulted in a conservative (but possibly very large) predicted CV for small estimates. Model 4 had the best overall RMSE, largely due to a downward curvature on the left side of the regression curve. Model 4 also resulted in a possible bias (understatement) of CV's for large estimates. (See Figures 1 through 5 for examples representative of the regression curve plots produced during the pilot test. See Figures 6 and 7 as examples where model 4 had lower RMSE than model 3 and Figures 7 and 8 as examples where model 4 had lower RMSE than model 3.)

Preliminary Conclusions

Models 2 and 5 were to be excluded from any further analysis based on the inappropriate shape of the regression curve fit to the data. More data would be needed for small estimates to choose between models 3 and 4. Model 1 would be included for further analysis because it is the only model with limited theoretical justification. It was therefore decided to fit all three viable models (models 1, 3 and 4) using three alternative fitting methodologies: unweighted, weighted, and iteratively reweighted non-linear regression approach.

b. <u>Exploratory Analysis</u>

Step 1: Percentages, totals and averages for selected variables from each of the four SASS data sets (School, School Administrator, Teacher, Teacher Demand & Shortage (TDS)) for various subdomains (i.e., region, state, school level, minority status, school size, community status and combinations of these) were calculated.

Step 2: CV's for the estimates in step 1 were calculated using balanced half-sample replication techniques. Plots of the log of the estimate versus the log of the CV were used to finalize groups to be used for model estimation.

Step 3: Using the calculated estimates in each of the subdomain groups from step 1 and their respective CV's from step 2, nonlinear models using SAS NLIN procedure were fit in order to produce ordinary least-squares (OLS), weighted least squares (WLS), and iteratively re-weighted least squares (IRLS) estimates of the parameters and respective R-squared values for each of the relvariance models 1, 3 and 4 described in section II. The WLS procedure was specified to work with the sum of squares which weighted inversely to the square of the observed CV and the IRLS method was specified to work with the sum of squares which weighted inversely to the square of the predicted CV. The minimizing values from the OLS technique were used as starting values in the WLS and IRLS runs. A plot of the regression curve fit for each of the three methods (OLS, WLS, IRLS) of fitting a model was used to determine which method for fitting the model worked best. Based on these plots, the IRLS technique of model fitting proved to be best. The OLS technique gave too much weight to the small estimates whose corresponding relvariance was usually large and unstable and the WLS technique was a better procedure because it gave the least reliable terms in the sum of the squares a reduced weight, but the IRLS technique fit most of the data better than either of the other two techniques. A plot showing the R^2 values of one model versus another model was used to determine which GVF model fit best. (See separate volumes for the above mentioned plots).

Step 4: An out of sample test was performed to validate conclusions made from step 3.

Findings: The following are the selected IRLS models within each survey based on the exploratory analysis:

-- **The School Survey**
Student Totals - GVF Model 3 was selected
Teacher Totals - GVF Model 3 was selected
Averages - GVF Model 1 was selected

-- **The TDS Survey**
Student Totals - GVF Model 1 was selected
Teacher Totals - GVF Model 1 was selected
Averages - GVF Model 3 was selected

-- **The School Administrator Survey**
Admin Percents - GVF Model 1 was selected
Admin Totals - GVF Model 1 was selected
Averages - GVF Model 3 was selected

-- **The Teacher Survey**
Teacher Percents - GVF Model 1 was selected
Teacher Totals - GVF Model 1 was selected

-- **Salary Averages**
- GVF Model 3 was selected

Standard Error of a Ratio

To estimate the relative variance of an estimated ratio, R = X/Y, where Y is an estimator of the total number of individuals in a certain subpopulation and X is an estimator of the number of individuals in another subpopulation, use

$$V^2_R = V^2_X - V^2_Y$$

where the relvariances of X and Y are read from the appropriate GVF table. This formula has been shown to produce useful approximations. The approximation is appropriate when the correlation between the ratio X/Y and the denominator Y is close to 0; the approximation is an overestimate if the correlation is positive.

IV. References

Kish, L. (1967). *Survey Sampling*. New York: John Wiley and Sons.

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

Figure 1

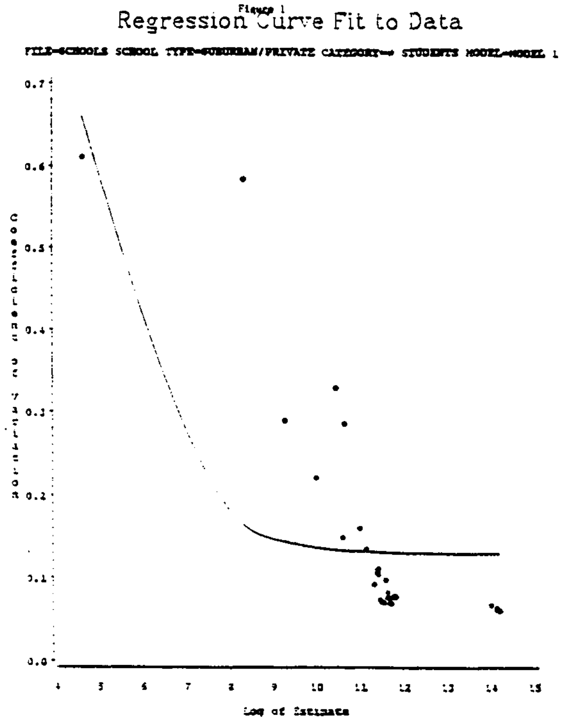


Figure 2

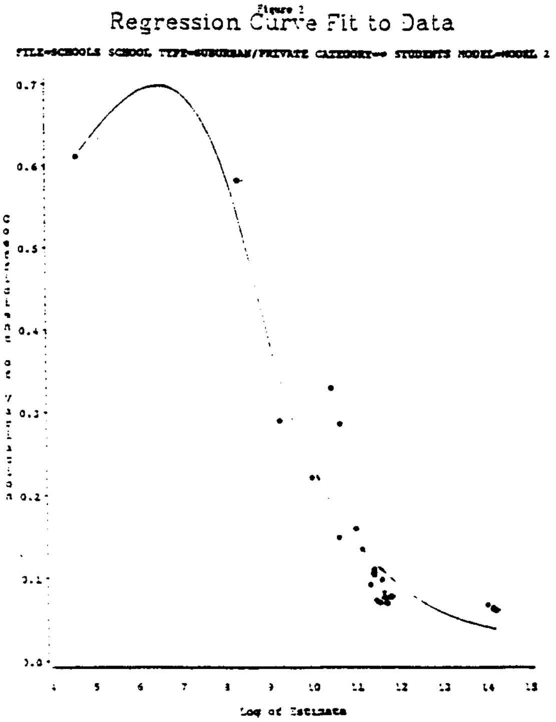


Figure 3

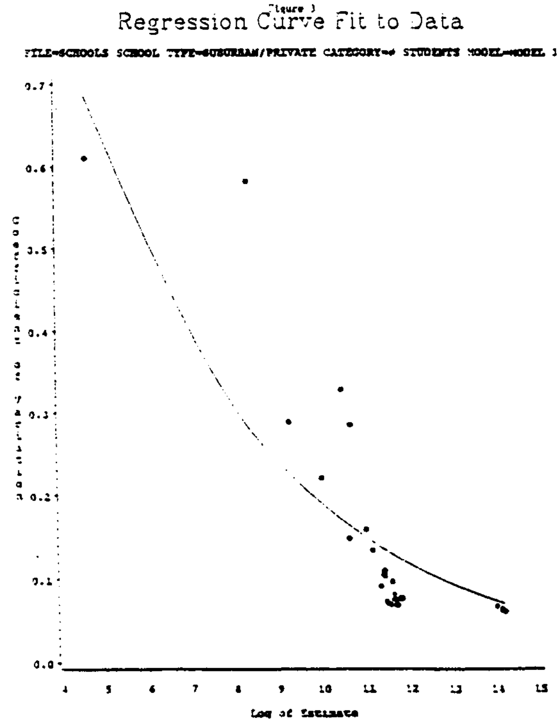


Figure 4

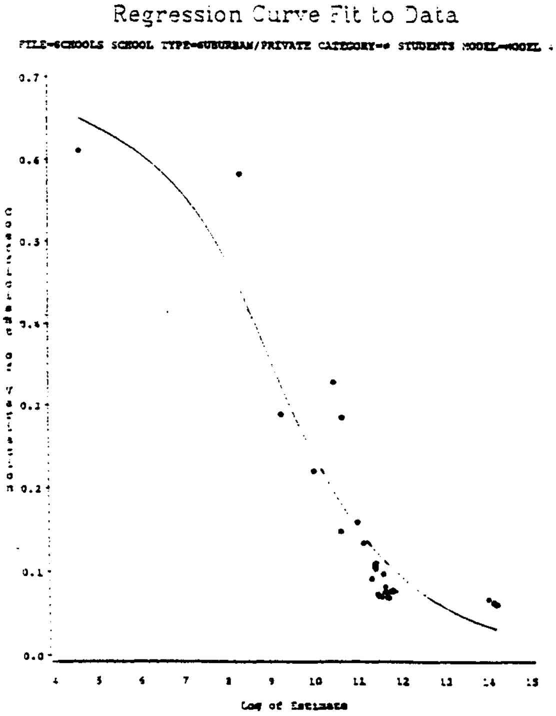


Figure 5

Regression Curve Fit to Data

FILE=SCHOOLS SCHOOL TYPE=SUBURBAN/PRIVATE CATEGORY=# STUDENTS MODEL=MODEL 5

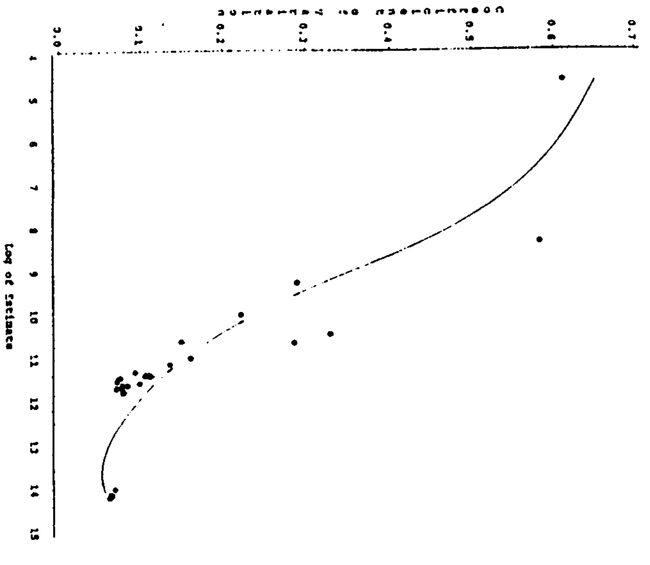


Figure 6

Regression Curve Fit to Data

FILE=SCHOOLS SCHOOL TYPE=RURAL/PUBLIC CATEGORY=# STUDENTS MODEL=MODEL 3

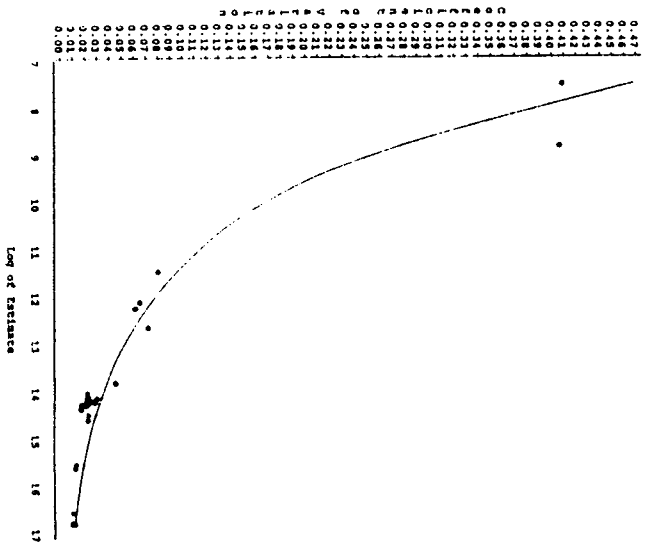


Figure 7

Regression Curve Fit to Data

FILE=SCHOOLS SCHOOL TYPE=RURAL/PUBLIC CATEGORY=# STUDENTS MODEL=MODEL 4

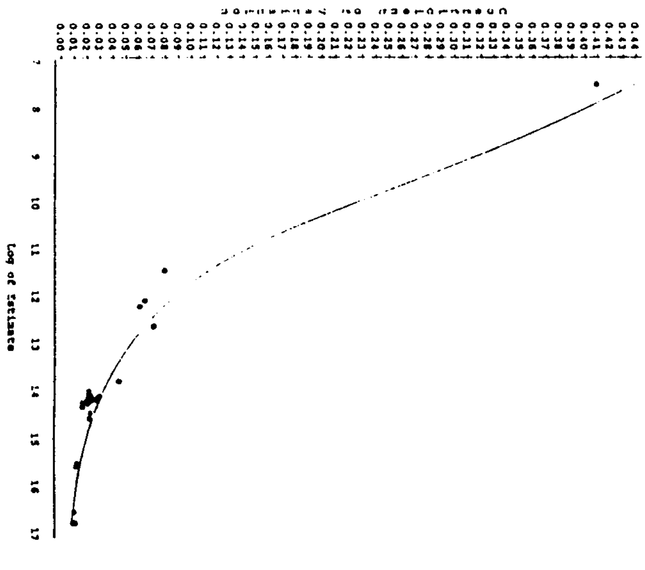


Figure 8

Regression Curve Fit to Data

FILE=SCHOOLS SCHOOL TYPE=SUBURBAN/PUBLIC CATEGORY=# STUDENTS MODEL=MODEL 3

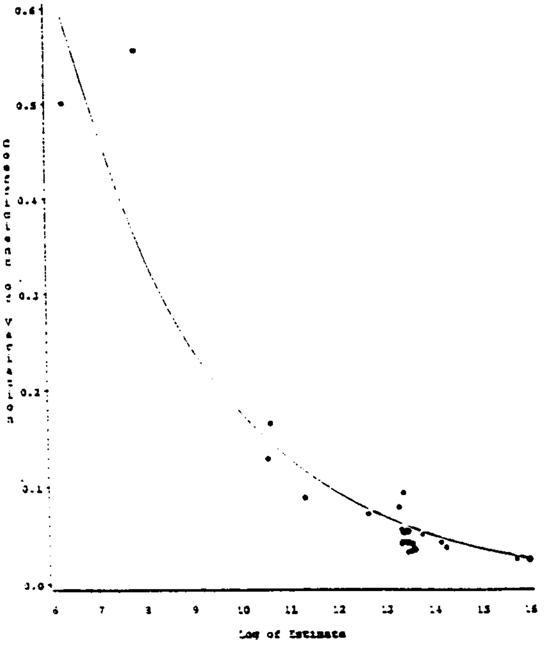
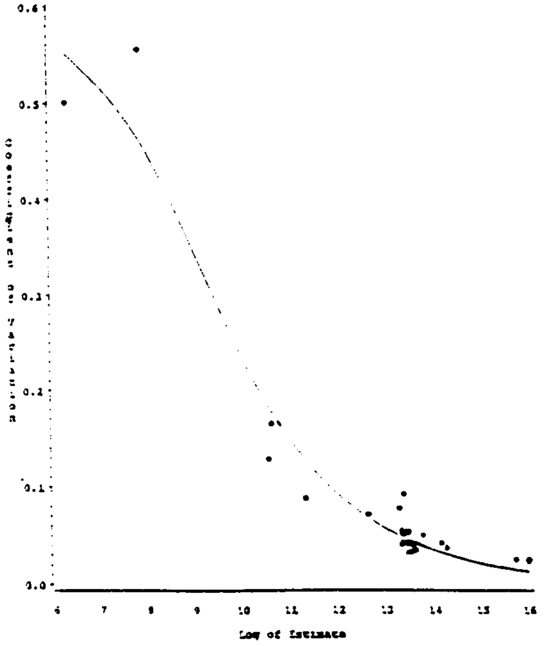


Figure 9

Regression Curve Fit to Data

FILE=SCHOOLS SCHOOL TYPE=SUBURBAN/PUBLIC CATEGORY=# STUDENTS MODEL=MODEL 4


Coefficient of Variation
Log of Estimate