GENERALIZED VARIANCE FUNCTIONS FOR DATA FROM MULTI-FRAME SURVEYS: THE SESTAT EXPERIENCE

Don S. Jang, Brenda G. Cox, and David J. Edson, Mathematica Policy Research, Inc.
Don S. Jang, MPR, 600 Maryland Ave. SW, Suite 550, Washington, DC 20024-2512

KEY WORDS: Variance approximation, GVF diagnostics, SESTAT

Abstract: Generalized variance functions (GVFs) provide a simple way to predict standard errors. For large multi-frame surveys, agencies usually estimate characteristics of interest for the individual surveys as well as for the combined survey. In this paper, we describe an investigation of GVF methods for such a multi-frame survey. The Scientist and Engineers Statistical Data System (SESTAT) combines information from three survey components. We begin by developing GVFs for each domain of interest for SESTAT as a whole. This global approach provides a simple and quick implementation of GVF methodology and produced relatively well predicted variance estimates. However, the global approach only gives standard errors for the combined SESTAT data set. Next, we pursued another approach to improve the GVF methodology. Each survey is conducted with different sampling designs, and has its own objectives. We combined the three survey-specific GVFs to create an integrated GVF, which can also be used to predict standard errors for the component surveys. The integrated GVF approach gave better results than the global approach, in the sense that it improved standard error prediction for most SESTAT domains. Moreover, given survey-specific GVFs for common domains of interest, standard errors for the combined data set can be predicted without any further statistical effort. However, the global model produced GVFs of acceptable quality and is simpler to use, leading to its adoption for SESTAT variance approximation.

1. Background

The Scientists and Engineers Statistical Data System (SESTAT) derives information on the labor force and education characteristics of U.S. scientists and engineers (S&E) from three component surveys: the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR). Each component survey is independently executed with its own sampling and analysis plan. SESTAT is a multi-frame survey in that it represents the entire S&E universe by merging these component surveys while properly accounting for multiple selection opportunities across surveys. For more details about multi-frame survey analysis, see Lessler and Kalsbeek 1992, Chapter 5.

The large number of data items in the SESTAT questionnaire makes it cumbersome to report standard errors for all statistics in survey reports. Instead of reporting individual standard errors for each estimate, SESTAT provides generalized variance functions (GVFs) so that users have a quick and simple way to calculate standard errors for survey estimates. GVFs predict the standard error based upon the relationship between a survey characteristic and its estimator's variance; the user inserts the estimated value of the characteristic of interest into the fitted GVF model to generate a model-based prediction of the variance.

In this paper, we discuss the methodology used to create GVFs for the 1993 SESTAT. We first review the customary GVF procedure and then describe how we developed GVFs following conventional methodology for SESTAT domains of special interest. This approach, which we refer to as the *global model*, provided a simple and quick implementation of GVF methodology. Next, we describe an approach to improve the GVF methodology for this multi-frame survey. The *integrated model* obtains the three survey-specific GVFs using the same general procedure used for the global model, and then integrates the survey-specific GVFs to create a GVF prediction model for the entire SESTAT. To assess the variance prediction ability of the global model versus the integrated model, we developed various diagnostic statistics. Our results indicate that the integrated GVFs are superior to the global GVFs, in that they provide more accurate standard error predictions for most SESTAT domains.

Some individuals in SESTAT's target population belong to the surveyed population of more than one component survey. For example, a bachelor at the time of the 1990 Census that went on to complete a master's degree in 1991 had opportunities for selection in the NSCG and the NSRCG. For both GVF models, we used a unique-linkage rule to remove multiple selection opportunities. Each member of SESTAT's target population was uniquely linked to one and only one component survey and then the individual was included in SESTAT only when selected for the linked survey.

Using the unique linkage rule, each person had only one chance of being included in the combined SESTAT database (see Carlson 1995).

2. The Global GVF Model

In this section, we describe how we developed GVFs for the combined SESTAT database by following the usual procedure for population totals (see, for instance, Wolter, 1985, pp.205-206).

2.1. Choosing the Set of Survey Variables to Use in Fitting the Model

The GVF approach begins with directly calculated variance estimates for a subset of all possible variables. Specification of appropriate variables to use in estimating the GVF model, then, is an important component of GVF procedures. A set of variables should be chosen that represents all variables of interest in the variance estimation sense. We chose 60 variables to use in fitting GVF models for 260 SESTAT domains of special interest.

2.2 Direct Variance Estimation Methods

Estimated totals and their design-based variance estimates are needed to fit the GVF model for each domain. Which variance estimation method is best for creating the design-based variance estimates depends on the specific sample design. Ultimately, the better direct variance estimators produce the better GVF models.

For the sake of simplicity, we assumed in this study that GVF model fit was not sensitive to the direct variance estimation method. Replicated variance estimation methods are needed for the SESTAT due to its complex sampling structure which makes deriving the usual design-based estimation formula difficult if not impossible. To compute variance estimators directly, we choose the random group method with 20 replications. The method of random groups draws multiple samples from a target population (or subpopulation) of interest and then constructs separate estimates for each replicate (e.g., Chapter 2, Wolter, 1985). The dispersion of the estimates across replicates provides the basis for the standard error measure.

2.3. Choosing the GVF Model Form

Many different forms of GVFs have been developed for use in approximating the variance of survey estimates. For population totals, GVF models are usually created for the relative variance of the estimated total $\hat{Y}$, or

$$RelVar(\hat{Y}) = \frac{Var(\hat{Y})}{Y^2}, \qquad (2.1)$$

where $Var(\hat{Y})$ is the variance of $\hat{Y}$. The modeling typically begins by assuming that the relative variance of the estimated total $\hat{Y}$ is a linear function of the inverse of the total Y being estimated, or

$$RelVar(\hat{Y}) = \beta_0 + \frac{\beta_1}{Y}. \qquad (2.2)$$

The parameters of the GVF model, β_0 and β_1, are unknown and estimated from a subset of all survey-derived totals and their variances by some form of regression estimation. Wolter (1985) provides the rationale for using model (2.2) but notes that there is little theoretical justification for any model. By simple linear transformation of (2.2), we obtain the GVF function used for the global model:

$$\mathrm{Var}(\hat{Y}) = \beta_0 Y^2 + \beta_1 Y. \qquad (2.3)$$

That is, it models the variance of the total estimates as a quadratic function of the totals.

2.4. Fitting Methodology

For simplicity and easy implementation, we employed the ordinary least squares method to fit the model in this exploratory investigation. (For production of actual GVF models, we recommend use of weighted least squares.) An estimator of the variance of an estimated total $\hat{Y}$ can be obtained by evaluating the GVF model at $\hat{Y}$ and at $\hat{\beta}_0$ and $\hat{\beta}_1$, which are the estimates of the GVF model parameters β_0 and β_1. Thus, using the GVF model, the standard error of a specific estimated total can be predicted by inserting the value of the estimated total into the following computational equivalent:

$$SE(\hat{Y}) = (\hat{\beta}_0 \hat{Y}^2 + \hat{\beta}_1 \hat{Y})^{1/2} \qquad (2.4)$$

where $SE(\hat{Y})$ is the predicted standard error of the estimated total $\hat{Y}$.

2.5. Report Model Parameter Estimates and Model R^2

After fitting the model, it is customary to report the model coefficient estimates and the values of R^2 (the percent of variation explained by the model). The model R^2 is a quick measure to judge the effectiveness of the GVF model for prediction of standard errors. The R^2 values can range from 0 to 1. If R^2 is close to unity, then the model is generally acceptable; values closer to 0 indicate that the model may be inaccurate. For the global model,

the R^2 values were mostly larger than 0.9. However, caution is needed in interpreting these values as a direct indication of good model fit because we used unscaled values for independent variables that have a large range, which could yield misleadingly large R^2 values.

3. The Integrated GVF Model

The integrated GVF model recognizes that a SESTAT estimated total is the simple sum of the three survey-estimated totals and (due to the use of unique linkage) the variance of the SESTAT total estimate is the simple sum of the associated survey-specific variances.

3.1. Survey-Specific GVFs with the Unique Linkage Cases

To produce the integrated GVF models, we first developed survey-specific GVFs for the three surveys by following the procedure described in Section 2, but with each SESTAT domain first partitioned into three survey-specific domains. Using unique-linkage, a SESTAT total is estimated by the sum of the survey-specific total estimates, or

$$\hat{Y}_{\text{SESTAT}} = \hat{Y}_{\text{NSCG}} + \hat{Y}_{\text{NSRCG}} + \hat{Y}_{\text{SDR}}. \qquad (3.1)$$

Due to the independence of three component surveys, the variance can also be estimated as the sum of the survey-specific variance estimates:

$$Var_{\text{SESTAT}} = Var_{\text{SDR}} + Var_{\text{NSCG}} + Var_{\text{NSRCG}}. \qquad (3.2)$$

As before, we used model (2.3) for all three survey-specific GVFs. Each of the 260 SESTAT domains then consists of three nonoverlapping survey-specific domains. It is thus possible to prepare GVFs for each survey for these 260 SESTAT domains.

For the 60 chosen variables, we calculated three survey-specific totals and their variances directly with the appropriate unique-linkage weight, using the same random group methodology as the global method. Instead of estimating an overall estimator $\hat{Y}$ and the corresponding variance of $\hat{Y}$, we produced estimates of survey-specific totals, $\hat{Y}_{\text{NSCG}}$, $\hat{Y}_{\text{SDR}}$, and $\hat{Y}_{\text{NSRCG}}$, and their variances, $\hat{V}_{\text{NSCG}}$, $\hat{V}_{\text{SDR}}$, and $\hat{V}_{\text{NSRCG}}$. For each domain, these survey-specific estimated totals and their variance estimates were used to fit survey-specific GVF models using an ordinary least squares method.

The R^2 values were quite high for most domains. Table 1 presents the resulting coefficient estimates for some survey-specific GVFs.

3.2. Integration of the Survey-Specific GVFs

The integrated model creates SESTAT standard error estimates based upon the three surveys' GVFs by recognizing that each survey uniquely describes one segment of target population with no overlap when the unique linkage approach is used. Assuming each survey produces good GVF models, the standard error estimates for SESTAT variables can be obtained from the integration of the three survey-specific GVFs using the following formula:

$$\hat{V}_{\text{Integrated}} = \hat{\beta}_{0,1}\hat{Y}^2_{\text{NSCG}} + \hat{\beta}_{1,1}\hat{Y}_{\text{NSCG}} + \hat{\beta}_{0,2}\hat{Y}^2_{\text{SDR}} + \hat{\beta}_{1,2}\hat{Y}_{\text{SDR}} + \hat{\beta}_{0,3}\hat{Y}^2_{\text{NSRCG}} + \hat{\beta}_{1,3}\hat{Y}_{\text{NSRCG}} \qquad (3.3)$$

Here, $\hat{\beta}_{0,1}$ and $\hat{\beta}_{1,1}$ are the estimated parameters for the NSCG subdomain of the particular domain of interest for SESTAT, and similarly for the other surveys. To obtain an approximate variance estimate of an estimated total using the integration of the three survey-specific GVFs, one uses the following procedures:

- The estimated totals for each of the three survey components using unique linkage cases are computed.
- The most appropriate SESTAT domain for the estimate is determined.
- Three sets of estimates of survey-specific parameters are obtained for this domain—one for each survey.
- The generalized variance is computed using equation (3.3).

For example, the estimate of the total number of scientists and engineers in 1993 is 11,615,174 with $\hat{Y}_{\text{NSCG}} = 10{,}129{,}126$; $\hat{Y}_{\text{SDR}} = 514{,}364$; and $\hat{Y}_{\text{NSRCG}} = 972{,}585$. To predict the variance, we get the three sets of survey-specific parameter estimates from the domain labeled "Total" in Table 1. Using the integrated GVF, an approximate standard error for the estimate of total scientists and engineers is:

$$(-0.000002\,\hat{Y}^2_{\text{NSCG}}+277.3\,\hat{Y}_{\text{NSCG}}+(-0.000027)\,\hat{Y}^2_{\text{SDR}}+$$
$$18.7\,\hat{Y}_{\text{SDR}}+0.003124\,\hat{Y}^2_{\text{NSRCG}}+234.6\,\hat{Y}_{\text{NSRCG}})^{1/2}=76{,}088.$$

The survey-specific GVFs can be used for the individual surveys as well as for integration.

4. Assess Global Vs. Integrated Models

To evaluate the global and integrated models, we used scatter plots and two numerical measures described below to evaluate the predictive abilities of the global and integrated models. These measures assess how successful each model is at describing the variation in the data. Note that there is no thorough theoretical criterion for the following methods due to lack of theoretical background for the GVF model. Instead, they give somewhat intuitive interpretations.

4.1. Scatter Plots

To get a feel for the performance of each model, it is useful to view, for each domain, graphs for the estimated and predicted variance of estimated totals versus the estimated total itself. We overlaid the plot of the fitted GVF curve onto the scatter plot of direct variance estimates, after a log-transformation to make the graphical presentation clearer. How well the fitted curve explains the observed variance estimates provided a visual demonstration of the goodness of fit. The scatter plots showed the GVF curves close to the actual standard errors for most domains, which indicates that both the global and integrated models give an intuitively reasonable approximation. We found the numerical measures easier to interpret and compare, however.

4.2. Absolute Relative Error

The first numeric measure looks at the error in the variance approximation. For this measure, we first calculated differences between the directly calculated standard error and the GVF-predicted standard error,

$$ERROR = SE_{\text{Actual}} - SE_{\text{Predicted}} \qquad (4.1)$$

where SE_{Actual} is the directly calculated standard error and $SE_{\text{Predicted}}$ is the predicted standard error from the GVF model. Next, we develop a scale-free measure of *ERROR* (expressed as a percentage),

$$REL\text{-}ERR = 100\,\frac{ERROR}{SE_{\text{Actual}}}\,. \qquad (4.2)$$

REL-ERR is a popular measure to check the adequacy of GVF standard errors, for example, it was used in the 1990-1991 School and Staffing Survey and the 1985 Young Adult Literacy Survey (Johnson and King, 1987). *REL-ERR* allows detection of patterns of underestimation or overestimation for GVF-based variance predictions. Moreover, the median of *REL-ERR* gives an indication of loss of accuracy when using GVF-derived standard errors.

The absolute value of *REL-ERR,*

$$AREL\text{-}ERR = |\,REL\text{-}ERR\,|\,, \qquad (4.3)$$

can be used to evaluate standard error predictions from the GVF models by counting the number of domains having *AREL-ERR* values greater than 20% (or some other set value). That is, *AREL-ERR* measures the relative loss of accuracy due to using GVF approximations. Because the standard errors computed by GVFs give an indication of the order of magnitude of the standard error of an estimate rather than the precise standard error, *AREL-ERR* can be a good measure to assess the adequacy of GVF approximations. *AREL-ERR* was used in developing the GVF models for the 1988 National Household Survey on Drug Abuse (Bieler and Williams, 1990). The average *AREL-ERR* for the set of estimates used to fit the GVF model measures the average distance between the actual versus predicted standard error, which we express as a percentage of the actual standard error. Small values for the average *AREL-ERR* indicate that the corresponding GVF fits well. Table 1 shows that the integrated approach yields improved predictability over the global approach for the 260 domains we used.

4.3. Absolute Error of the Coefficient of Variation

We also propose two other numerical measures,

$$CV\text{-}DIF = 100\,\frac{ERROR}{Y} \qquad (4.4)$$

and

$$ACV\text{-}DIF = |\,CV\text{-}DIF\,|. \qquad (4.5)$$

To see if a large deviation of a GVF predicted standard error from the actual standard error matters in a practical sense, *ACV-DIF* values should be evaluated. For example, the GVF model for a specific domain having an average *AREL-ERR* greater than 20% will lead to a different width for the confidence interval from that of the direct variance estimator. However, if the

corresponding *ACV-DIF* is less than 5%, then the resulting interval estimation in a practical sense might not be sensitive to the difference of confidence interval width, because the difference is relatively small in comparison to the total being estimated. In summary, the measure *ACV-DIF* indicates whether in a practical sense the GVF model for a specific domain having an average *AREL-ERR* greater than 20% can lead to significantly different interval estimation from that using a direct variance estimator. Roughly, we might say that if the *ACV-DIF* values are all less than 5%, then *AREL-ERR*s larger than 20% may be tolerable. The comparison between the global and integrated models using the *ACV-DIF* measure produced similar results to those for *AREL-ERR*. The *ACV-DIF* values for the global and integrated models indicate that the integrated model offers enhanced prediction capabilities.

5. Summary

Table 1 shows the results of the generalized variance procedure and the performance of the global and integrated models, respectively. The first column identifies the domain of interest. The third and fourth columns contain the parameter estimates for the intercept and the slope. The next two columns list the diagnostic measures for checking the GVFs —*AREL-ERR* and *ACV-DIF*, respectively. The integrated model is more suitable than the global model when these numerical measures are compared. Moreover, using the three survey-specific GVFs, users can produce standard errors for SESTAT as a whole as well as for the three individual surveys. However, the integrated model is more complex and requires that the user know the values of the three survey-specific estimates as well as that of the combined total, making it impractical for use with project reports. On the other hand, the global approach produces variance estimates of acceptable quality (though not as good as the integrated approach) and is much simpler to use and explain. For these reasons, SESTAT has adopted the global approach for use in variance approximation. Full documentation for approximating SESTAT standard error and results of the investigation are found in its own home page (http://srsstats.sbe.nsf.gov/stderr00.html).

REFERENCES

Bieler, G. S., and R. L. Williams. "Generalized Standard Error Models for Proportions in Complex Design Surveys." *Proceedings of Section on Survey Research Methods of the American Statistical Association,* 1990, pp. 272-277.

Carlson, Barbara Lepidus. "Coverage and Multiplicity Issues in the SESTAT Data System." Report to the National Science Foundation, February, 1995.

Cox, Brenda G., Don S. Jang, and David Edson. "Sampling Errors for SESTAT and Its Component Surveys." Report to the National Science Foundation, December, 1996.

Jang, Don. S., and Brenda G. Cox. "Survey-Specific GVFs for The 1993 SESTAT." Report to the National Science Foundation, November, 1996.

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

Lessler, Judith T., and William D. Kalsbeek. *Nonsampling Error in Surveys.* New York: John Wiley & Sons, 1992.

National Center for Education Statistics. *Design Effects and Generalized Variance Functions for the 1990-91 Schools and Staffing Surveys (SASS).* Vol I, II. Washington, DC: U.S. Department of Education, 1995.

Valliant, Richard. "Generalized Variance Functions in Stratified Two-Stage Sampling." *Journal of the American Statistical Association*, Vol 82, 1987, pp. 499-508.

Wolter, Kirk M. *Introduction to Variance Estimation.* New York: Springer-Verlag, 1985.

TABLE 1

RESULTS OF THE GENERALIZED VARIANCE MODELING FOR THE 1993 SESTAT

Domain	GVF		Intercept	Slope	AREL-ERR (%)	ACV-DIF (%)
Total S&Es	Model 1		0.000029	176.70	16	0
	Model 2	NSCG	-0.000002	277.33	16	0
		SDR	-0.000027	18.71		
		NSRCG	0.003124	234.56		
Male S&Es	Model 1		0.000031	166.31	14	0
	Model 2	NSCG	-0.000004	257.37	15	0
		SDR	-0.000033	19.49		
		NSRCG	0.002513	195.90		
Female S&Es	Model 1		-0.000008	270.33	15	1
	Model 2	NSCG	0.000018	234.31	15	1
		SDR	-0.000085	11.66		
		NSRCG	0.004559	147.63		
White S&Es	Model 1		0.000029	201.00	17	0
	Model 2	NSCG	-0.000005	317.79	18	0
		SDR	-0.000031	20.11		
		NSRCG	0.003977	136.21		
Nonwhite S&Es	Model 1		0.000185	83.19	12	1
	Model 2	NSCG	0.000004	114.79	15	1
		SDR	0.000094	15.04		
		NSRCG	0.011837	295.17		
Bachelor S&Es	Model 1		0.000088	118.86	21	1
	Model 2	NSCG	0.000011	253.52	13	0
		NSRCG	0.004198	198.80		
Masters S&Es	Model 1		-0.000042	210.07	15	1
	Model 2	NSCG	-0.000043	208.35	15	1
		NSRCG	0.001538	139.09		
Doctorate S&Es	Model 1		-0.000028	66.04	18	1
	Model 2	NSCG	-0.000197	172.36	16	1
		SDR	-0.000027	18.71		
		NSRCG	0.004268	30.27		