Generalized Variance Functions for the 2003 Survey of Doctorate Recipients

Y. Michael Yang[1], Yongyi Wang[2]
National Opinion Research Center, 1350 Connecticut Ave., NW, Washington D.C. 20036[1]
National Opinion Research Center, 55 E. Monroe St., Chicago, IL 60603[2]

Abstract

A generalized variance function (GVF) is a mathematical model describing the relationship between the variance or relative variance of a survey estimator and its expectation. The Survey of Doctorate Recipients (SDR) has been publishing GVF parameters for major analysis domains of interest since the early 1990s. This paper compares the 2003 SDR variance estimates derived from several common GVF models. The purpose is to evaluate the existing GVF model and search for a potentially superior model than the simple linear model that has been used. The predicted variance from each model is compared with the directly estimated variance and a model is considered superior if the predicted variance is closer to the direct variance estimate.

Keywords: SDR, GVF, Variance estimation

1. Background

The Survey of Doctorate Recipients (SDR) is a longitudinal survey of individuals who have received doctorate degrees from U.S. institutions in science and engineering (S&E) fields and are pursuing their professional careers in the United States. The SDR design is complex due to the numerous redesigns since its inception in 1973. The 2003 SDR may be characterized as a stratified random design where stratification is defined by degree field, gender, and demographic group. The SDR sample is selected systematically within each stratum. To account for the complex design, the Successive Difference Replication (SUD) method was used to compute the variance of SDR statistics (Wolter, 1984; Fay and Train, 1995; Tupek, 2003). GVFs are considered for SDR because it is not feasible to directly calculate and publish the variance for all SDR statistics. In particular, it is impossible to anticipate the various analysis domains that may be of interest to SDR data users. The GVFs provide a mechanism for the data users to compute the variance of their estimates that are not directly provided by SDR.

2. GVF Steps

A GVF is a mathematical model describing the relationship between the variance or relative variance of a survey estimator and its expectation. The SDR has been publishing GVF parameters for major analysis domains or population subgroups since the early 1990s. To conduct variance estimation using GVF method, the following steps were implemented.

First, to account for potential differences across important population subgroups, the GVFs needed to be estimated for each subgroup independently. We defined 352 subgroups by crossing 32 degree field groups and 11 demographic groups. As analysis domains, these subgroups are not mutually exclusive. Some subgroups represent a subset of other subgroups. For example, female Computer and Information Sciences doctorate recipients is a subset of all female doctorate recipients in Sciences, which is a subset of all doctorate recipients. For subgroups that are not covered by this classification, the analyst may use the GVF estimated for all doctorate recipients combined, i.e., Total Doctorate Recipients/Total Population.

In the second step, the set of key SDR variables to be used in direct variance estimation was identified. These variables have been determined to be important analysis variables. In addition, they are sufficiently diverse in the sense that the observed totals cover a wide range of values within each analysis domain. Once the set of key variables were identified, they were used to define 103 SDR statistics. All these statistics were estimated totals.

The third step computed the direct point and variance estimates for the 103 statistics for each population subgroup using the successive difference replication method. SUDAAN's DESCRIPT procedure was used to carry out these estimates with the replication method of Balanced Repeated Replication (BRR). The finite population correction (FPC) factor was applied to all direct variance estimates outside SUDAAN.

The final step was to fit the GVF models using the direct point and variance estimates from the previous step as input. Parameters derived from these GVF models would then allow the data users to approximate the variances for SDR statistics that are not directly estimated under SUD.

3. GVF Models

There are several mathematical models that can be used as generalized variance functions to describe the relationship between the variance of a survey estimate and its expectation. Most of the models are based on the assumption that the relative variance is a decreasing function of the magnitude of the mean or expectation.

3.1 The 2003 SDR GVF Model

The GVF model used for the 2003 SDR is as follows:

$$V^2 = a + b/X \qquad [1]$$

where $V^2 = Var(\hat{X})/X^2$ denotes the relative variance, $\hat{X}$ is an estimator of the total number of cases possessing some characteristics, $X = E(\hat{X})$ is the expectation of $\hat{X}$, $Var(\hat{X})$ is the variance of $\hat{X}$, and a and b are the generalized variance function parameters to be estimated.

For each of the 352 population subgroups, the GVF parameters a and b were estimated through an iterative weighted linear regression procedure using the direct point and variance estimates as input. The purpose of using weighted linear regression is to improve the reliability of the fitted model by assigning relatively smaller weights to less reliable direct variance estimates and larger weights to more reliable direct variance estimates. The iterative weighted linear regression procedure involves four steps at each iteration. The regression weight at each step for model 1 is described in Table 1.

The initial regression weight at step 1 is the inverse of the squared relative variance. In the subsequent steps, the regression weight is replaced by the inverse of the squared relative variance that is estimated from the previous step. At the end of step 4, observations with an absolute standardized residual exceeding 3 were identified as outliers and were removed from further consideration. After that, the second iteration starts and the four-step regression procedure is repeated on the remaining observations. This iterative process continues until all absolute standardized residuals are smaller than 3. At the conclusion of the regression procedure, we obtain the estimated parameters *a* and *b* from the final model.

3.2 Four other GVF Models

We now consider the four additional models below.

$$V^2 = a + b/X + c/X^2 \qquad [2]$$

$$V^2 = (a + bX)^{-1} \qquad [3]$$

$$V^2 = (a + bX + cX^2)^{-1} \qquad [4]$$

$$\log(V^2) = a - b\log(X^2) \qquad [5]$$

All these models are discussed in Wolter (1985). Our interest is to see if any of these models would perform better than Model 1. A model is considered better if the model-predicted variance is closer to the directly estimated variance. For each of the four models, we estimated the model parameters using the same iterative weighted linear regression procedure as applied to estimating model 1. The regression models and their regression weights at each step for Models 2-5 are summarized in Tables 2-5.

4. Model Evaluation

To evaluate the five models, we divided the 103 SDR statistics into two sets: an estimation set and a validation set. Statistics in the estimation set were used to estimate the GVF parameters for each of the five models and for each of the 352 domains based on the procedures described above. The estimated parameters were then used to predict the relative variance for the statistics in the validation set for each of the domains. These GVF predicted relative variances were then compared with the relative variances that were estimated "directly" under the SUD method.

To form the estimation and validation sets, we first sorted the 103 SDR statistics by the magnitude of the direct variance estimate. Then we selected a systematic sample of 68 statistics to form the estimation set and used the remaining 35 statistics as the validation set. The purpose of sorting was to ensure that estimates of various magnitudes were represented in both groups.

For each GVF model and population subgroup, we estimated the GVF parameters based on the direct estimates of the 68 statistics in the estimation set. These estimated GVF parameters were then used to predict the relative variance for the 35 statistics in the validation set. Thus, for each of the 35 statistics in the validation set, two variance estimates were available: one from direct estimation under SUD and one from the GVF estimation. This pair of estimated variances formed the basis for our evaluation. We call the estimates from SUD the "direct variances" or "actual variances" and the estimates from GVF the "predicted variances." The performance of each GVF model was to be evaluated by comparing the predicted variances

with the actual variances. A GVF model was considered superior if on average its predicted variances were closer to the direct variances.

For each of the 35 statistics in the validation set, we obtained a pair of predicted and actual variances for each of the 352 population subgroups with at least 20 cases. The smallest subgroups were dropped because their estimates would be less reliable. In addition, we deleted data points where the point estimate was 0. This left us 8,967 pairs of predicted and actual variances under each GVF model, representing various statistics by domain combinations (there were initially 35*352=12,320 possible statistics by domain combinations). For each pair of estimated variances, we computed the difference of the predicted variance from the actual variance. Then, for each GVF model, we computed the mean difference over all pairs.

Table 6 reports the mean and standard deviation of the relative differences between the predicted and the actual variances under each model. The relative difference is defined as the difference between the predicted variance and the actual variance divided by the actual variance, with the result expressed as a percentage. Table 6 shows that, relative to the actual variances, all but one models tend to overestimate the variance on average. Under Model 1, the predicted variances are on average 17.2 percent higher than the actual variances. For Model 2, the predicted variances are on average 15.4 percent higher. Similarly, Model 5 overestimates the variances by 36.6 percent on average. Model 4 is almost on target on average but the relative differences have large variations. Model 3, an obvious aberration here, by far has the least predictive power.

Table 7 compares the predictive power of the five models in terms of absolute relative differences between the predicted variance and the actual variance. The calculation of absolute differences ignored the direction of the differences. So for Model 1, the predicted variance is 35.4 percent off the actual variance on average. For Model 2, the predicted variances are off by 34.6 percent. Again, Models 1 and 2 perform much better than the other three models.

To alleviate the effect of possible outliers on the mean relative differences, we further divided the relative differences between the predicted and actual variances into five categories by the magnitude of the relative differences. Table 8 shows the cumulative percentages at four levels of relative difference. For Model 1, 17.6 percent of the differences between the predicted and the actual variances are within 5 percent of the actual variances; 35.2 percent of the differences between the predicted and the actual variances are within 10 percent of the actual variances; 63.5 percent of the differences between the predicted and the actual variances are within 20 percent of the actual variances; and 79.4 percent of the differences between the predicted and the actual variances are within 30 percent of the actual variances. The reported percentages under the other models should be interpreted in the same manner. For example, under Model 2, 81 percent of the differences between the predicted and the actual variances are within 30 percent of the actual variances.

Table 8 depicts a more favorable picture about the performance of the GVF models, especially for Models 1 and 2. It is reassuring that the vast majority of the GVF predicted variances are quite close to the directly estimated variances under Model 1. However, this cannot be said of the other three models. For Model 3, only about a third of the relative differences are within 30 percent of the actual variances; and for Models 4 and 5, slightly over half of the differences are within 30 percent of the actual variances.

5. Discussion

GVF estimation has been a major component of variance estimation for the SDR. The purpose of our study is to evaluate the existing GVF model and possibly identify a superior model for future GVF estimation. In general, the GVF model used for the 2003 SDR worked well in terms of approximating the direct variance estimates. Across a large number of domains and SDR statistics, the overwhelming majority of the predicted variances are quite close to the variance estimates derived from the SUD method. In particular, the average difference between the predicted and the actual variances is positive, indicating overestimation which is generally considered less problematic than underestimation.

Our evaluation has shown consistently that Model 2 may perform better than the existing model although the potential improvement is likely to be small. A closer look at the two models may be necessary before a formal recommendation can be made about future GVF modeling for the SDR. The performance of the other three models is far from satisfactory and these models may be dropped from further consideration.

We finally note a few limitations of the study. First, we focused on comparing the average performance of these models while in fact the models may perform differently for different statistics and domains. For example, if most of the outliers are associated with certain domains or statistics, the GVF models may be

improved by taking that information into account. Second, there might be a better model for the SDR beyond the five models evaluated here. We plan to look into this possibility later.

References

Fay, R. E. and Train, G. F. (1995), "Aspects of Survey and Model-based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," ASA Proceedings of the Section on Government Statistics, 154-159.

Tupek, A.R. (2003). "Calculation of Generalized Variance Parameters for the 2001 Survey of Doctorate Recipients (SDR01-VAR-3)," Internal Census Bureau Memorandum, February 11, 2003.

Wolter, K.M. (1984). "An Investigation of Some Estimators of Variance for Systematic Sampling." Journal of the American Statistical Association, vol. 79, no. 388, pp. 781-790.

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

Table 1. Regression Model and Weight at Each Step for GVF Model 1: $V^2 = a + b/X$

Step	Model	Weight
1	$\frac{Var(\hat{X})}{X^2} = a_1 + \frac{b_1}{X}$	$\left[\frac{X^2}{Var(X)}\right]^2$
2	$\frac{Var(\hat{X})}{X^2} = a_2 + \frac{b_2}{X}$	$\frac{1}{\left[a_1 + \frac{b_1}{X}\right]^2}$
3	$\frac{Var(\hat{X})}{X^2} = a_3 + \frac{b_3}{X}$	$\frac{1}{\left[a_2 + \frac{b_2}{X}\right]^2}$
4	$\frac{Var(\hat{X})}{X^2} = a_4 + \frac{b_4}{X}$	$\frac{1}{\left[a_3 + \frac{b_3}{X}\right]^2}$

Table 2. Regression Model and Weight at Each Step for GVF Model 2: $V^2 = a + b/X + c/X^2$

Step	Model	Weight
1	$\frac{Var(\hat{X})}{X^2} = a_1 + \frac{b_1}{X} + \frac{c_1}{X^2}$	$\left[\frac{X^2}{Var(X)}\right]^2$
2	$\frac{Var(\hat{X})}{X^2} = a_2 + \frac{b_2}{X} + \frac{c_2}{X^2}$	$\frac{1}{\left[a_1 + \frac{b_1}{X} + \frac{c_1}{X^2}\right]^2}$
3	$\frac{Var(\hat{X})}{X^2} = a_3 + \frac{b_3}{X} + \frac{c_3}{X^2}$	$\frac{1}{\left[a_2 + \frac{b_2}{X} + \frac{c_2}{X^2}\right]^2}$
4	$\frac{Var(\hat{X})}{X^2} = a_4 + \frac{b_4}{X} + \frac{c_4}{X^2}$	$\frac{1}{\left[a_3 + \frac{b_3}{X} + \frac{c_3}{X^2}\right]^2}$

Table 3. Regression Model and Weight at Each Step for GVF Model 3: $V^2 = (a + bX)^{-1}$

Step	Model	Weight
1	$\frac{X^2}{Var(\hat{X})} = a_1 + b_1 X$	$\left[\frac{X^2}{Var(X)}\right]^2$
2	$\frac{X^2}{Var(\hat{X})} = a_2 + b_2 X$	$(a_1 + b_1 X)^2$
3	$\frac{X^2}{Var(\hat{X})} = a_3 + b_3 X$	$(a_2 + b_2 X)^2$
4	$\frac{X^2}{Var(\hat{X})} = a_4 + b_4 X$	$(a_3 + b_3 X)^2$

Table 4. Regression Model and Weight at Each Step for GVF Model 4: $V^2 = (a + bX + cX^2)^{-1}$

Step	Model	Weight
1	$\frac{X^2}{Var(\hat{X})} = a_1 + b_1 X + c_1 X^2$	$\left[\frac{X^2}{Var(X)}\right]^2$
2	$\frac{X^2}{Var(\hat{X})} = a_2 + b_2 X + c_2 X^2$	$(a_1 + b_1 X + c_1 X^2)^2$
3	$\frac{X^2}{Var(\hat{X})} = a_3 + b_3 X + c_3 X^2$	$(a_2 + b_2 X + c_2 X^2)^2$
4	$\frac{X^2}{Var(\hat{X})} = a_4 + b_4 X + c_4 X^2$	$(a_3 + b_3 X + c_3 X^2)^2$

Table 5. Regression Model and Weight at Each Step for GVF Model 5: $\log(V^2) = a - b\log(X)$

Step	Model	Weight
1	$\log\left(\frac{Var(\hat{X})}{X^2}\right) = a_1 - b_1\log(X)$	$\frac{1}{\left[\log\left(\frac{Var(X)}{X^2}\right)\right]^2}$
2	$\log\left(\frac{Var(\hat{X})}{X^2}\right) = a_2 - b_2\log(X)$	$\frac{1}{[a_1 - b_1\log(X)]^2}$
3	$\log\left(\frac{Var(\hat{X})}{X^2}\right) = a_3 - b_3\log(X)$	$\frac{1}{[a_2 - b_2\log(X)]^2}$
4	$\log\left(\frac{Var(\hat{X})}{X^2}\right) = a_4 - b_4\log(X)$	$\frac{1}{[a_3 - b_3\log(X)]^2}$

Table 6. Mean Relative Difference of Predicted Variances from Actual Variances (n=8,967)

model	Mean (%)	Std Dev (%)
1	17.2	102.3
2	15.4	134.0
3	5974.6	569948.3
4	-0.2	1231.4
5	36.6	137.4

Table 7. Mean Absolute Relative Difference of Predicted Variances from Actual Variances (n=8,967)

model	Mean (%)	Std Dev (%)
1	35.4	97.6
2	34.6	130.4
3	6189.7	569946.0
4	99.0	1227.4
5	62.9	127.6

Table 8. Cumulative Percentages at Different Levels of Relative Difference

Deviation	Model 1	Model 2	Model 3	Model 4	Model 5
<= 5%	17.6	18.5	7.1	13.3	8.5
<=10%	35.2	36.3	13.1	24.0	17.4
<=20%	63.5	65.3	24.1	41.3	35.5
<=30%	79.4	81.0	33.6	52.8	52.6