Journal of Official Statistics
Vol. 3. No. 3, 1987, pp. 235 – 250

Generalized Variance Functions for a Complex Sample Survey

Eugene G. Johnson and Benjamin F. King[1]

Abstracts: For a national survey of reading ability among young adults using a multistage, stratified probability sample, generalized variance functions (GVFs) are estimated. That is, an attempt is made to express the estimated variance of a statistic as a function of that statistic and other characteristics of the variable of interest. With GVFs estimated from a development sample of variables, predictions of sampling variance are made for other variables in a confirmation sample and comparisons made with conventional jackknife estimates. Conclusions are drawn about the feasibility of use of GVFs, with emphasis on the margin of additional estimation error that is introduced.

Key words; Estimated variance; design effects; jackknife.

1. Introduction

In this paper we report the results of an investigation into the feasibility of using generalized variance functions (GVFs) for estimation of sampling variances for statistics computed for a large-scale and complex survey. As described in Wolter(1985), the GVF method attempts to model the variance of a survey estimator as a function of the estimate and possibly other variables. If the modeling is successful then it is unnecessary to compute the estimated variance by the usual formula thus accruing considerable cost savings. An accurate GVF may also be of great value in designing similar surveys in the future. This approach to variance estimation has been adopted by the Bureau of the Census for the Current Population Survey, and also by the National Center for Health Services Research in certain applications. In addition to the references cited in Wolter(1985), the reader should see Cohen(1979), Cohen and Kalsbeek(1981), and Burt and Cohen(1984). Valliant(1987), in a paper just published at this writing, provides interesting theoretical justification in terms of a prediction model for the use of GVFs.

In many of the previous applications the GVF is used to model the relative variance of an estimated subpopulation total. In the traditional approach it is assumed that the relative variance of an estimated total, $\hat{X}$, is a decreasing function of the magnitude of the estimate. A common specification is

$$\text{relvar}\,(\hat{X}) = \alpha + \beta/\hat{X}. \qquad (1.1)$$

This specification is in turn used to derive a model for the relative variance of a ratio or proportion. In this paper we focus primarily on the direct estimation of variances of proportions, e.g., for the percentage of subjects who choose a particular answer in an achieve-

[1] Survey Methods, Research Statistics Group, Educational Testing Service, Princeton, NJ 08541, U.S.A.

ment test. We shall also devote some discussion to the modeling of variances of subpopulation totals.

The statistics of interest are from the Young Adult Literacy Survey, conducted in the summer of 1985 by Response Analysis Corporation for the Educational Testing Service Center for National Assessment of Educational Progress. The target population was all persons of age 21 through 25 residing in households in the Continental U.S. The sample design involved five stages of selection with stratification at several of the stages. The units of selection by stage were: (1) counties, groups of counties, or metropolitan statistical areas (MSAs); (2) census tracts, groups of tracts, or segments of tracts; (3) blocks of contiguous housing units within tracts; (4) housing units within blocks; and (5) those eligible within housing units. The 25 largest MSAs were included in the sample with certainty, and the remaining noncertainty MSAs and counties ordered geographically within separate strata. A systematic selection of 65 primary sampling units from these strata was made with probability proportional to size (pps). A total of 400 second-stage units (SSUs) was drawn with pps from these primary units and from the 25 selfrepresenting MSAs. In this stage of selection SSUs with high densities of Blacks or Hispanics were oversampled at an approximate rate of 2 to 1 to permit special focus in the study on those groups. Two blocks were selected (pps) from each of the 400 SSUs, and 48 housing units were chosen within each block. A total of 38 400 housing units were screened for eligible subjects, resulting in a final sample of 3 618 respondents, each of whom provided measures of cognitive and general background characteristics. Estimates of means, totals, and proportions obtained for these items involve weights that reflect adjustments for disproportionate sampling, nonresponse, and poststratification to known marginal totals.

This type of design will produce variances different from the variances produced by simple random sampling (srs) with fixed sample size. This is so for a number of reasons. There are gains in precision over that of srs from stratification by geography and size. These gains, however, are counterbalanced by the effects of nonoptimal disproportionate selection and clustering. The weights, themselves, are subject to sampling variability which makes the statistics of interest nonlinear. All of these considerations combine to make the estimators of sampling variances more complex and computationally more expensive than the easy srs algorithms.

In the Young Adult Literacy Survey the variance estimation procedure was the *jackknife*, see, e.g., Wolter (1985), p.185. The certainty MSAs and the remaining primary units yielded a total of 98 clusters which for the purpose of variance estimation were successively paired throughout the frame. Members of the resulting 49 pairs are from the same size strata and the same geographic area. The jackknife technique involves the computation of forty-nine pseudo-values of each statistic of interest by successively omitting one member of each pair.

The purpose of our research is to develop an alternative estimator of sampling variability that demands less computation, but is of adequate precision. The general approach is to fit linear models of functions of the sampling variance to estimates from the survey, using the jackknife variance estimate as the basis for the dependent variable, and various easily-computed statistics as the predictors. For items not used in the development of the model, variances are estimated by prediction from the fitted equation. For large-scale surveys, a relatively small number of items would be used in GVF development thus avoiding the computation costs of conventional variance estimation for the remaining

variables. In repeated execution of surveys of the same population and with the same types of variables, it might even be possible to use parameter estimates for the prediction model from earlier applications.

In Section 2 of this paper we discuss two proposed criteria for the measurement of the goodness of prediction of the models that we shall examine. The first is a measure of absolute error and the second a measure of performance in terms of relative error. Then in Section 3 we describe our various attempts at modeling the variances of estimated proportions (ratios). In Section 4 the results of the model fitting are evaluated in terms of a "fundamental model" in which the sample variance is expressed as a product of a systematic factor (i.e., the true variance) and a random noise factor. In Section 5 we show the gain in predictive ability from prior knowledge of the design effect. There is a brief discussion in Section 6 of the results from model fitting for the variances of estimated domain totals, and finally in Section 7 some conclusions are stated.

2. Criteria for Goodness of Fit

Our aim in developing a GVF is to predict the variance of a statistic for use in estimation and inference. For this study, and clearly for many other studies, underestimation of sampling variability is a more serious error than overestimation. Thus we would rather have estimated standard errors that are too large than those that are too small. For illustration in this paper, we shall assume that the consequences of an underestimate are *three* times as severe as those of an overestimate of the same magnitude. We shall also assume the opportunity loss to be linear. A standard result from decision theory, e.g., Raiffa and Schlaifer (1961) shows that the predicted value of the dependent variable that minimizes expected linear opportunity loss of the error of estimation is the quantile of the predictive distribution given by the ratio

$$k_u/(k_u + k_o),$$

where k_u and k_o are the losses of under and overestimation, respectively. Hence, assuming normality, we use the following expression as the optimal prediction for the various models.

$$Y^* = \text{predicted value given by the model} + .67 \text{ standard error of prediction}, \tag{2.1}$$

corresponding to the .75 quantile of the normal probability distribution. We shall evaluate alternative models by comparing the means of *absolute residuals* from the optimal predicted values, Y^*, weighting positive residuals by three. After we have discussed some of the alternative forms of the GVF in the following section, we shall show that the results of using this approach to model evaluation are robust against the choice of k_u and k_o, the losses of under and overestimation.

To understand more fully the implications of the models it is useful to consider goodness of prediction from a different angle. We have stated that we are primarily concerned with *underestimation* of the jackknife standard error, i.e., we would prefer estimates that are too big, rather than too small. For model comparision we plan to give underestimates three times the weight of overestimates and to compute the weighted mean absolute error (average loss). It will be seen, however, that the best models in our application are logarithmic and it seems reasonable to ask for some measure of performance in terms of the antilogs, i.e., the standard error that we are ultimately interested in knowing. We will therefore transform the predicted values,

Y^*, to predictions of the standard error and compute a relative error of prediction:

$$\text{relative error} = \frac{\text{(jackknife s.e.} - \text{predicted s.e.)}}{\text{jackknife s.e.}}, \tag{2.2}$$

thus expressing the under or overestimate as a proportion of the jackknife standard error. We shall then be able to answer questions such as "What fraction of the estimates from the model are downward biased by more than 20 percent?" We believe that reports of model performance in that form are often more interesting than the comparison of R-squares, mean square errors, or average losses.

3. Variance Models for Cognitive Items

In the Young Adult Literacy Survey there were 104 different items, each designed to measure a person's cognitive ability with respect to one of four psychometric scales: (1) reading proficiency, (2) prose comprehension, (3) document utilization, and (4) practical computation. The principal statistics for any one of these items are the proportions of subjects choosing each of a set of possible response categories. Statistics for each item were produced for each of 13 domains of the target population. These domains include the total population, each sex, three racial groups, three levels of education, and four geographic regions. Thus we have measures on 104×13 or 1 352 variables for which we have arbitrarily chosen the first response category for this analysis. The items are sufficiently diverse so that observed proportions cover a wide range within each domain. For model development we have further drawn a systematic sample of 897 of the item measures, saving the remainder for validation. The selection was balanced to provide the same 69 items for each of the 13 domains. In the survey, items were administered according to a balanced-incomplete-block spiraling scheme so that not every person was given each item. The average number of responses to an item for the total population was 1 487, and obviously smaller for narrower domains; for example, for Hispanics the average number of cases was only 167.

3.1. *The traditional model for proportions*

The first GVF model for proportions that we consider is derived from the model for the estimated total, shown in (1.1) above. The derivation is based on the assumption that equation (1.1) holds for both numerator and denominator of the sample ratio (proportion) and that there is zero correlation between the ratio and its denominator (Wolter (1985, p.204)).

Let $\hat{A}$ be the estimated total in a subpopulation possessing a certain attribute, $\hat{X}$ be the estimated total size of the subpopulation, and define $\hat{p} = \hat{A}/\hat{X}$ to be the estimated sample proportion.

Then we write

$$\text{relvar}\,(\hat{p}) = \beta\,(1-\hat{p})/\hat{p}\hat{X}\,, \tag{3.1}$$

or in terms of the variance,

$$\text{var}\,(\hat{p}) = \beta\hat{p}\,(1-\hat{p})/\hat{X}\,. \tag{3.2}$$

Exhibit 1A shows the results of an ordinary least squares fit of the model in (3.2) above, with the intercept allowed to be nonzero.

Exhibit 1A. Unweighted Least Squares Regression for Var ($\hat{p}$): Model (3.2)

Predictor variables	Coefficient	STD error	Student's *t*
Constant	1.86	.18	10.08
$\hat{p}(1-\hat{p})/\hat{X}$	6449	161.20	40.02

Degrees of freedom	895
R—square	.64
Residual mean square	17.79

Residuals vs. Predicted Values

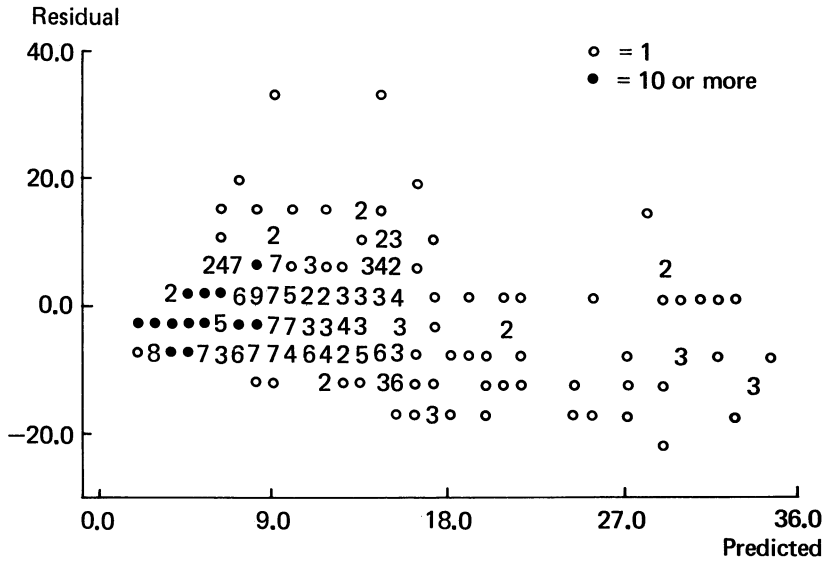
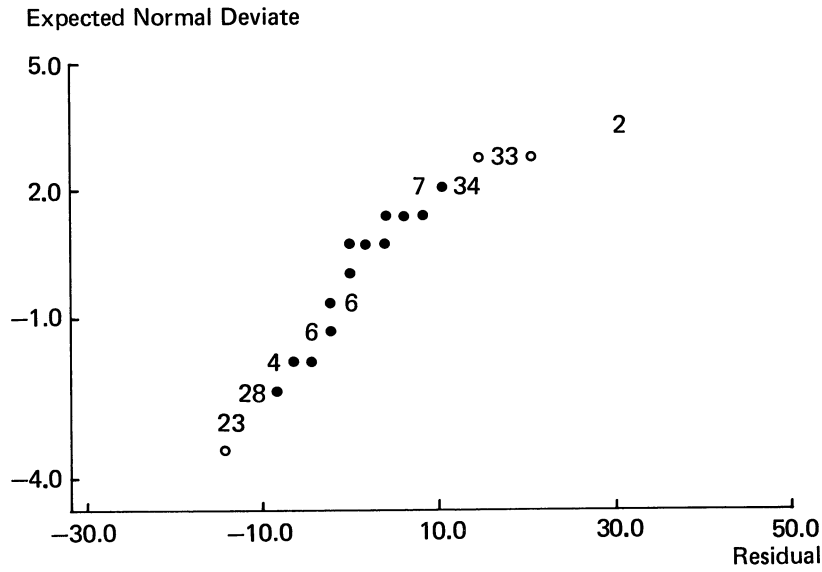


Exhibit 1B. Normal Probability Plot of Residuals from Model (3.2)



A noteworthy aspect is the asymmetry about zero in the plot of residuals against predicted values. The nonlinearity of the normal probability plot, Exhibit 1B, demonstrates further violation of the usual regression assumptions. The large residual mean squared error indicates poor predictability for the dependent variable.

Exhibit 2 shows the results of a regression of the variance of $\hat{p}$ on the simple random sampling formulation,

$$\text{var}(\hat{p}) = \alpha + \beta\hat{p}(1-\hat{p})/n\,, \qquad (3.3)$$

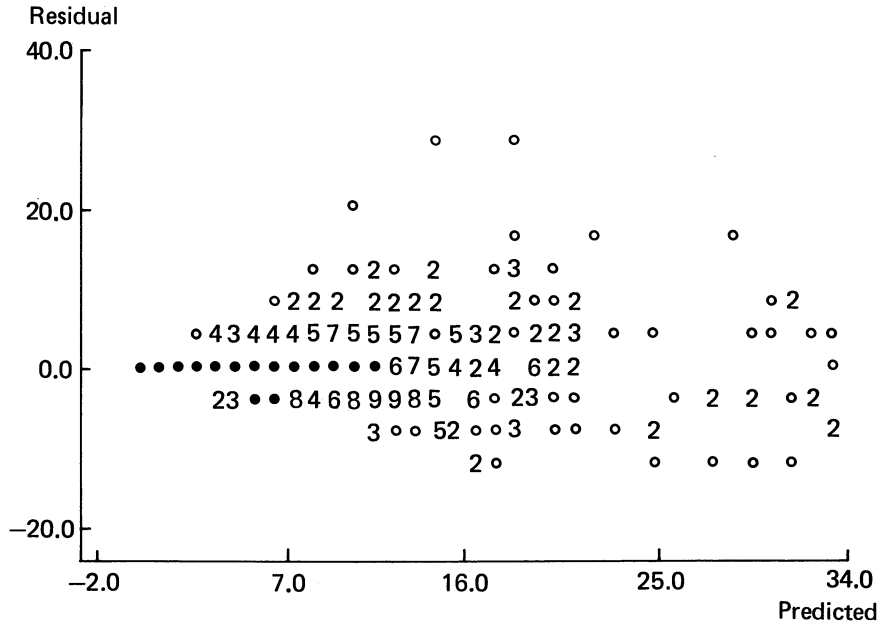
where n is the sample size. It can be seen that the linear fit is better than that for (3.2) above, leading us to doubt the effectiveness of the traditional specification. The residuals, however, still exhibit asymmetry.

Exhibit 2. Unweighted Least Squares Regression for Var (p): Model (3.3)

Predictor variables	Coefficient	STD error	Student's *t*
Constant	−.36	.17	−2.10
$\hat{p}(1-\hat{p})/\hat{n}$	2.09	.04	53.35

Degrees of freedom	895
R−square	.76
Residual mean square	11.87

Residuals vs. Predicted Values



3.2. Transformation to logarithms

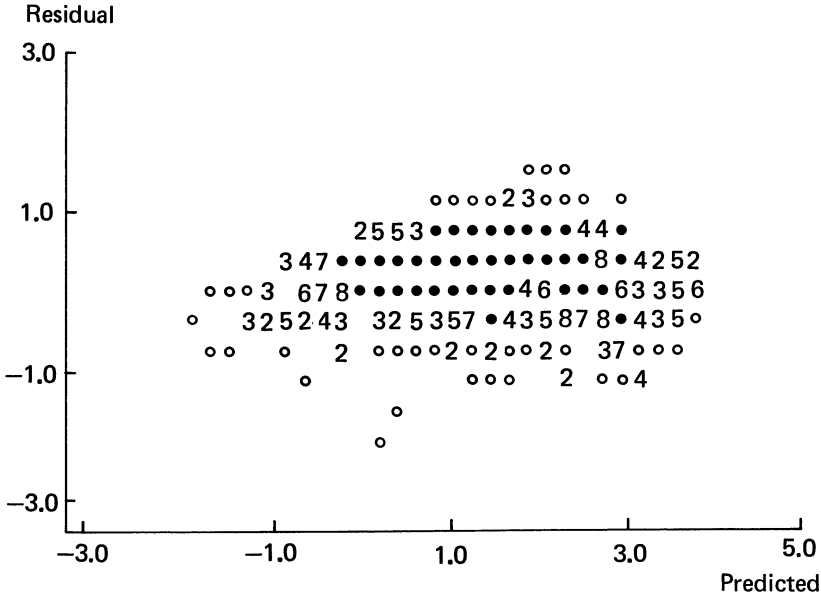
One of the problems with the approach taken thus far is that the model parameters have been estimated by conventional least squares, the optimality of which depends on underlying normality. Since sample variances have skewed distributions, one should perhaps not expect symmetry in residuals. There is also discussion in the GVF literature of the necessity to correct for inconstant residual variances, for example using iterative weighted least squares or nonlinear techniques. The alternative approach used here is to transform to logarithms, fitting linear models by least squares. This transformation makes the errors more symmetric, more homoskedastic, and reduces the impact of extreme values. An additional advantage is that the transformation to logarithms converts multiplicative relationships to linear relationships so that the models for variance, relative variance, standard error, and design effect all

Exhibit 3A. Unweighted Least Squares Regression for Logvar ($\hat{p}$): Model (3.4)

Predictor variables	Coefficient	STD error	Student's t
Constant	8.45	.11	79.33
$\text{Log}(\hat{p}(1-\hat{p})/\hat{X})$	.91	.01	67.41

Degrees of freedom	895
R—square	.84
Residual mean square	.21

Residuals vs. Predicted Values



have similar forms. One should expect the advantages of the logarithmic transformation to carry over to other GVF applications.

As an example, we transform the variables in the model in (3.2) to their natural logarithms, and express their relationship by

$$\text{logvar}\,(\hat{p}) = \alpha + \beta \log\,(\hat{p}\,(1-\hat{p})/\hat{X}). \quad (3.4)$$

The results of ordinary least squares are shown in Exhibits 3A and 3B. The scatter plot and the normal probability plot show that the residuals from this regression follow the normal distribution more closely than in earlier specifications. For this reason we shall continue to work in the logarithmic metric.

In the following display we show the various forms of logarithmic models for the variance of $\hat{p}$ that we have investigated. In addition, we report the value of R^2, the residual mean square, and the average loss using $k_u = 3$ and $k_o = 1$, discussed in Section 2.

Exhibit 3B. Normal Probability Plot of Residuals from Model (3.4)

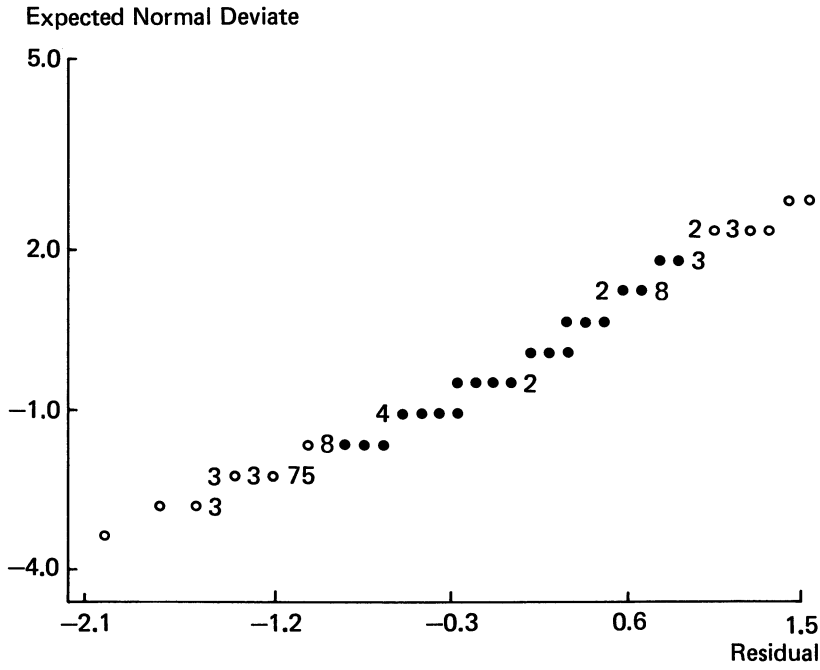


Exhibit 4. Logarithmic Models for the Variance of $\hat{p}$

$$\text{logvar}(\hat{p}) = \alpha + \beta \log(\hat{p}(1-\hat{p})/\hat{X})$$
$R^2 = .84 \quad RMS = .21 \quad AVLOSS = .52$
Model (3.4)

$$\text{logvar}(\hat{p}) = \alpha + \beta \log(\hat{p}(1-\hat{p})/n)$$
$R^2 = .89 \quad RMS = .14 \quad AVLOSS = .44$
Model (3.5)

$$\text{logvar}(\hat{p}) = \alpha + \beta_1 \log(\hat{p}) + \beta_2 \log(1-\hat{p}) + \beta_3 \log(n)$$
$R^2 = .90 \quad RMS = .13 \quad AVLOSS = .44$
Model (3.6)

$$\text{logvar}(\hat{p}) = \alpha + \beta_1 \log(\hat{p}) + \beta_2 \log(1-\hat{p}) + \beta_3 \log(n) + \beta_4 \log(\text{cv}(\hat{X})),$$
$R^2 = .90 \quad RMS = .13 \quad AVLOSS = .43$
Model (3.7)

$$\text{logvar}(\hat{p}) = \alpha + \beta_1 \log(\hat{p}) + \beta_2 \log(1-\hat{p}) + \beta_3 \log(n) + \beta_4 G3 + \beta_5 G4 + \beta_6 G5 + \beta_7 G8 + \beta_8 G9 + \beta_9 G10$$
$R^2 = .91 \quad RMS = .12 \quad AVLOSS = .42$
Model (3.8)

Discussion:

Recall that Model (3.3) gave a better fit than Model (3.2). Since Models (3.5) and (3.4) are logarithmic versions of (3.3) and (3.2), respectively, we are not surprised to see that (3.5) outperforms its predecessor (3.4). Model (3.6) allows separate coefficients for the factors in the simple random sampling variance for the proportion, but its improvement over (3.5) is only slight.

Model (3.7) introduces the logarithm of the coefficient of variation of the estimated population total into the equation in the hope of taking account of unequal cluster sizes and unequal weights. Although the estimated regression coefficient of the new variable is significant, the increment to R^2 and the reduction in *AVLOSS* is negligible.

Finally, Model (3.8) is motivated by an interest in generalizing the results of the GVFs to other surveys and statistics. An important question is whether the parameters of the models examined thus far are constant across different classes of items and different population subgroups. Although scatter plots of residuals did not indicate great variability from group to group, we experimented with the introduction of class (domain) effects and developed the best fitting Model (3.8). The new variables $G3$, $G4$, $G5$, $G8$, $G9$, and $G10$ are one-zero indicators for the domains female, whites, blacks, high school education, greater than high school, and Northeast Region, respectively. In the above display we see again that the increase in R^2 is not great, and the decrease in average loss is about two percent from that of Model (3.7).

3.3. *The mean design effect model*

The last model for proportions that we consider is obtained by constraining ß to be equal to one in Model (3.5), i.e.,

$$\text{logvar}\,(\hat{p}) = \alpha + \log\,(\hat{p}\,(1-\hat{p})/n)\,, \quad (3.9)$$

for which least squares yields the *mean log design effect* as the estimator of α. Thus the sample logvar is expressed as the mean log design effect plus the log of the estimator from simple random sampling. The average logdeff for the set of 897 items is .55 and the standard deviation is .39, implying an *RMS* of about .15. The calculation of *AVLOSS* yields .46. Therefore the performance of Model (3.9) is not much worse than that of Model (3.6), and its simplicity makes it appealing. Somewhat discouraged by the lack of improvement from the introduction of identifiers for domains of study and classes of items, we decided to stop the data dredging and conclude that there is little hope of developing a GVF model that can perform a great deal better than (3.9).

3.4. *Robustness of the mean loss criterion*

To check on the robustness of the mean loss criterion we compare Models (3.4) and (3.6) for various values of k_u with k_o set at one. The results, shown in Table 1, demonstrate that the average loss for each model is approximately linear in Z, where Z is the $k_u/(k_u + k_o)$ quantile of the unit normal distribution. We also see that the superiority of Model (3.6) is maintained throughout. It is our conjecture that this model's superiority for all k_u cannot be proved analytically, but that there are few instances where Model (3.6) is not superior to Model (3.4).

3.5. *Relative error of estimation*

We next consider the second performance criterion discussed in Section 2 given by

Table 1. Comparison of AVLOSS for Models (3.4) and (3.6)

k_o	k_u	Z-value	*AVLOSS* (3.4)	*AVLOSS* (3.6)
1	1	.00	.35	.27
1	2	.44	.46	.37
1	3	.67	.52	.44
1	4	.84	.57	.48
1	8	1.22	.69	.59
1	20	1.67	.85	.73

Expression (2.2). For Model (3.9) we display in Exhibit 5 the histogram for the relative errors computed as shown in (2.2). It can be seen that 218 out of the 897 errors (24.3 percent) are positive, that is, the standard error is underestimated. The relative error of underestimation does not exceed 40 percent. The histogram also shows that only 26 out of 897 (2.9 percent) are underestimates greater than 20 percent in relative terms. The median among the errors of underestimation is less than 10 percent. The maximum relative error of overestimation is 176.7 percent, i.e., the predicted standard error was 2.77 times larger than the jackknife estimate. It can be seen, however, that such extreme overpredictions are rare. In only 70 cases (7.8 percent) did the predicted value of the standard error exceed the jackknife estimate by more than 50 percent. Finally, we observe that out of the total of 897 relative errors, 555 or 61.9 percent were less than 20 percent in absolute value.

Exhibit 5. Distribution of Relative Error (Model 3.9)

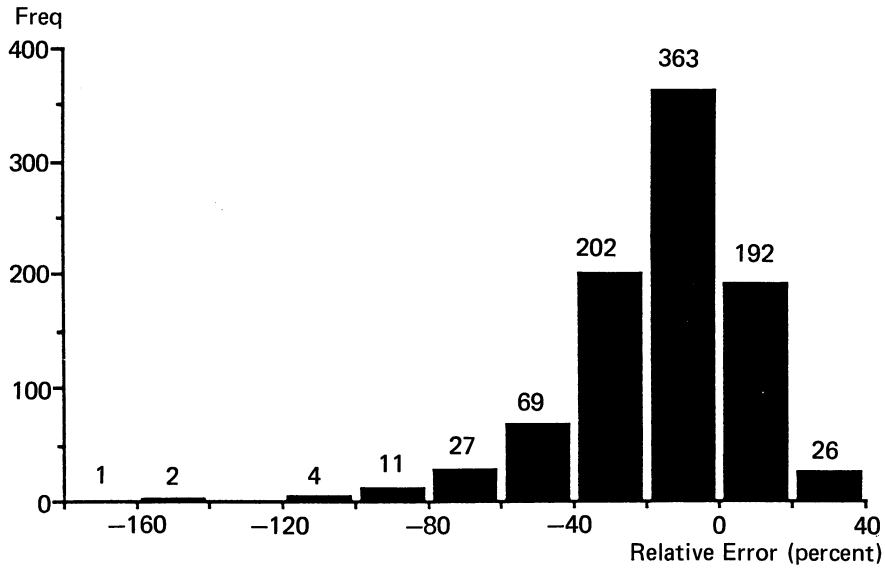
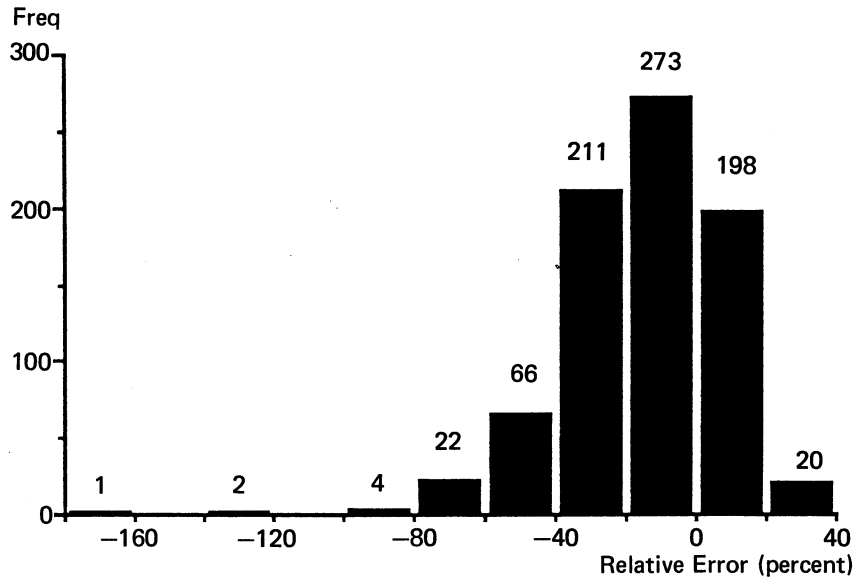


Exhibit 6 shows the relative errors for Model (3.6) above. The maximum relative error of underestimation is now only 35 percent. In only 20 cases (2.2 percent) did the relative error of underestimation exceed 20 percent. Out of the total of 897, 571 (63.7 percent) lie between minus and plus 20 percent.

One might ask whether the logarithmic transformation that we have used starting with Model (3.4) is all that necessary for good prediction of the standard error of the estimate. We have calculated the relative errors for the traditional Model (3.2), based on the empirical results shown in Exhibit (3.1). For the prediction equation in this exercise, using a version of Expression (2.1), we use

Exhibit 6. Distribution of Relative Error (Model 3.6)



var* =

predicted value + .25 std. error of pred.,

where, since the residuals are not normal, the coefficient .25 has been empirically determined to correspond to the 75th quantile. Table 2 shows a comparison of relative errors of prediction for Model (3.2) and the logarithmic Models (3.6) and (3.9):

Table 2. Distribution of Relative Errors by Model

Range of error	Relative frequencies in range of error		
	Model (3.2)	Model (3.6)	Model (3.9)
−20% underpred.	.07	.02	.03
20% − 0% underpred.	.18	.22	.21
0% − 20% overpred.	.22	.42	.41
20% − 50% overpred.	.29	.29	.27
50% − overpred.	.24	.06	.08
	1.00	1.00	1.00

The most striking result in the table is the serious overprediction of Model (3.2) in comparison to the logarithmic models. Nearly one-fourth of the standard errors are overpredicted by more than 50 percent. Note also that the frequency of errors of underprediction by more than 20 percent is greather than twice that of the other two models. In contrast to roughly two-thirds for the logarithmic models, only 40 percent of the relative errors for Model (3.2) are within 20 percent in absolute value.

3.6. Validation

The prediction equations estimated for Models (3.6) and (3.9) were applied to the 455 cases that had been held in reserve for validation. For Model (3.9) the average loss in the validation run was .50, as opposed to .46 in the original fitting. Curiously, only 16.7 percent of the relative errors computed according to Expression (2.2) were underestimates. The maximum relative underestimate was 34.6 percent. For Model (3.6), the average loss for the validation sample was .48, as opposed to .44 originally. The errors of underestimation constituted 21.1 percent of the total, with a maximum value of 39.2 percent. The validation run confirms our earlier finding that there is no great advantage in prediction accuracy of the more complicated Model (3.6) over the simple approach in Model (3.9) of adding the average logdeff to the logarithm of $\hat{p}\hat{q}/n$.

4. A "Fundamental Model" for the Variance of the Estimator

We should not be too critical of GVF estimation of standard errors without considering the performance that can reasonably be expected from even the most well-fitting model. The estimate of the standard error is itself subject to sampling error. Letting sj_i^2 be the jackknife variance estimator for variable i, we write

$$sj_i^2 = \sigma_i^2 \varepsilon_i ,$$

where ε_i is approximately distributed as a chi-square random variable divided by d degrees of freedom. It follows that

$$\log (sj_i^2) = \log (\sigma_i^2) + \log (\varepsilon_i) .$$

In our efforts to fit a model, we have assumed that there is some "ideal" linear relationship of the form

$$\log (\sigma_i^2) = \alpha + \beta x_i + \delta_i ,$$

which implies the following model for the jackknife variance estimator:

$$\log(sj_i^2) = \alpha + \beta x_i + (\delta_i + \log(\varepsilon_i)).$$

In this model, x_i is a predictor that can be extended to several predictors for the variable of interest if necessary, and δ_i represents the remaining unspecified but systematic sources of variation in the prediction of $\log(\sigma_i^2)$. The remaining term, $\log(\varepsilon_i)$, is the noise in estimation of $\log(\sigma_i^2)$.

It can be shown (e.g., Scheffe (1959, p.84)) that the variance of this noise is approximately $2/d$, where d represents the degrees of freedom. The degrees of freedom for a jackknife estimate from our study can be no larger than 49, and will generally be smaller. Thus $2/d$ ranges from a minimum of about .04 to a maximum of 2.00. For an intermediate number of degrees of freedom, say 17, the value is .12. This figure corresponds to the value of the residual mean squared error for our best fitting GVF equation, as shown in Exhibit 4. Thus it may well be that we have succeeded in explaining the systematic component of variability in logvar and that all that remains is noise[2].

5. Using Prior Knowledge of the Design Effect

As a final exercise with the cognitive items we shall introduce the design effect into the right-hand side of the model. To do so exactly would lead to a perfect fit by tautology. To be a bit more realistic, we assume that the analyst has only a rough prior idea of the magnitude of the effect. In his discussion of theoretical motivations for Models (1.1) and (3.1), Wolter(1985) suggests that the specification is consistent with a constant deff for groups of items. The logarithm of the deff ranges from –1.23 to 1.81 in these data.

[2] We are indebted to our colleague Paul Holland for discussions that led to these observations on the goodness of fit of our models.

Assume that it is possible a priori to place a proportion in one of the four categories of logdeff: (1) less than – .9, (2) greater than or equal to – .9 and less than zero, (3) greater than or equal to zero and less than +.9, (4) greater than or equal to +.9. With the variables $LD1$ and $LD2$ and $LD3$ as indicators, the item falls into one of the first three of the logdeff categories above. We can then fit the following model:

$$\begin{aligned}\text{logvar}(\hat{p}) = \alpha &+ \beta_1 \log(\hat{p}) + \beta_2 \log(1-\hat{p}) \\ &+ \beta_3 \log(n) + \beta_4 LD1 \\ &+ \beta_5 LD2 + \beta_6 LD3. \quad (5.1)\end{aligned}$$

R^2 is .97 and the residual mean square is .04. The mean loss with the 3:1 penalty for underestimation that we have been using falls to .26, only 62 percent of the previous minimum. This superiority is further borne out in the examination of relative errors of estimation of the standard error. The range is – 41.7 to 24.4 percent, with 777 out of 897 (86.6 percent) of the relative errors less than 20 percent in absolute value.

To be even more realistic we assume that the best that the analyst can do *a priori* is to place the proportion in the two categories: negative vs. nonnegative logdeff. In other words, deff is less than, greater than, or equal to one. In the following model $LDSIGN$ is a zero-one indicator of the sign of the log design effect:

$$\begin{aligned}\text{logvar}(\hat{p}) = \alpha &+ \beta_1 \log(\hat{p}) + \beta_2 \log(1-\hat{p}) \\ &+ \beta_3 \log(n) + \beta_4 LDSIGN. \quad (5.2)\end{aligned}$$

This fit is still better than those in models that do not involve logdeff, with $R^2 = .93$ and $RMS = .09$. The mean loss of estimation error is calculated to be .38. The range of relative error of estimation of the standard error is – 74.8 to 36.2 percent with about two-thirds falling within plus or minus 20 percent.

6. Estimation of Domain Totals

In addition to the cognitive items, for which we have been discussing GVFs for variances of proportions, the Young Adult Literacy Survey provides information on 214 background items, and covers the 13 different subpopulations of interest. Thus there is a large number of weighted estimates of domain totals with their corresponding jackknife estimates of variance. A systematic sample of 947 estimated totals (72 or 73 values for each of the 13 domains) was selected for analysis, with many other values held in reserve for subsequent exploration and validation.

As in the case of variances of proportions, the traditional Model (1.1) for predicting relvar $(\hat{X})$ and the model for var $(\hat{X})$ derived from it performed very poorly, and their residuals were badly skewed and heteroskedastic. Hence we transformed to logarithms and, after considerable fishing, determined the following two best specifications:

$$\text{logvar}(\hat{X}) = \alpha + \beta_1 \log(\hat{X}) + \beta_2 \log^2(\hat{X})$$
$$R^2 = .94 \quad RMS = .35 \quad AVLOSS = .81$$
Model (6.1)

$$\text{logvar}(\hat{X}) = \alpha + \beta_1 \log(\hat{X}) + \beta_2 L3 + \beta_3 L4 + \beta_4 L5 + \beta_5 L7 + \beta_6 L9 + \beta_7 L10 + \beta_8 L11 + \beta_9 L12 + \beta_{10} L13$$
$$R^2 = .98 \quad RMS = .13 \quad AVLOSS = .42$$
Model (6.2)

The more complex Model (6.2) was motivated by the hope that the introduction of effects for the thirteen principal domains in the population would greatly increase the predictive power. As is well known, it is impossible to design a survey that provides equally efficient estimates for all populations of interest. Thus we would expect design effects to vary from domain to domain, and accordingly, the parameters of the GVF to be different from group to group. We examined a fully specified model with intercepts and interactions for all groups. As expected, the fit of the model was markedly better than earlier specifications. After eliminating certain group effects that did not appear to be significant, we arrived at the parsimonious and effective Model (6.2), where the Ls are interactions between log $(\hat{X})$ and various domain indicators, and the corresponding βs are the incremental partial regression coefficients for those interactions. The domains used in the model are, in order, (3) females, (4) whites, (5) blacks, (7) less than high school education, (9) greater than high school, (10) Northeast region, (11) Southeast, (12) Central, and (13) West. The fit of Model (6.2) is only trivially different from that of the specification with separate intercepts and interactions for all thirteen domains.

The following table displays the relative distribution for the two models:

Table 3. Distribution of Relative Errors

Relative frequencies in range of error		
Range of error	Model (6.1)	Model (6.2)
–20% underpred.	.07	.04
20% under-overpred.	.32	.68
20%–overpred.	.61	.28
	1.00	1.00

It can be seen that Model (6.2) has a considerably tighter distribution of relative errors.

We made a validation run for Model (6.2) applying the prediction equation to 947 unused cases. The average loss for these cases was .47, and the maximum relative error of underestimation was 57.9 percent. The number of relative errors between minus 20 percent and plus 20 percent was 609, or 64.3 percent of the total cases. The percentage of relative errors exceeding 20 percent (i.e. underestimation) was only 4.6.

In summary, our attempts at modeling the variance of the total are no more or less successful than those for the variance of a proportion. By some criteria the results may be deemed adequate, but it is still possible to have very large errors in using the GVF to estimate jackknife values. It would be interesting to see similar measures of performance (e.g., distributions of relative errors) for other major applications such as the Current Population Survey of the U.S. Bureau of the Census.

7. Conclusion

We have investigated a sequence of models for predicting the variances of estimated proportions and totals. Our investigations have shown that prediction is improved by transforming to the logarithmic scale. We also found that in the Young Adult Literacy Survey the relatively simple model for proportions based on the average log design effect works nearly as well as more complex specifications and that its prediction error is about at the level of the noise of the jackknife estimator. As we have shown, the only way to markedly improve upon the simple model is by using prior information about the design effects of individual estimators.

In the Young Adult Literacy Survey the GVF was not used since our research took place after publication of the final report. The cost reduction from using the generalized variance function instead of resampling methods of estimation depends on the size of the data base, the number of variables for which variance estimates are required, and, most important, the number of replicates. We estimate that application of a GVF in the present case would have saved at least 90 percent of the cost of calculation of means and variances.

8. References

Burt, V.L. and Cohen, S.B. (1984): A Comparison of Methods to Approximate Standard Errors for Complex Survey Data. Review of Public Data Use, 12, pp. 159–168.

Cohen, S.B. (1979): An Assessment of Curve Smoothing Strategies which Yield Variance Estimates from Complex Survey Data. American Statistical Association, Proceedings of the Survey Research Methods Section.

Cohen, S.B. and Kalsbeek, W.D. (1981): NMCES Estimation and Sampling Variances in the Household Survey. National Center for Health Services Research.

Raiffa, H. and Schlaifer, R. (1961): Applied Statistical Decision Theory. Harvard University, Boston.

Scheffe, H. (1959): The Analysis of Variance. John Wiley & Sons, New York.

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

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

Received January 1987
Revised September 1987