GENERALIZED STANDARD ERROR MODELS FOR PROPORTIONS IN COMPLEX DESIGN SURVEYS

Gayle S. Bieler and Rick L. Williams, Research Triangle Institue
Gayle S. Bieler, PO Box 12194, Research Triangle Park, NC 27709-2194

1. Introduction

Generalized variance functions are often employed when large numbers of estimates are to be published from a survey. Generalized variance functions lessen the volume of published reports where presentation of each standard error estimate would essentially double the size of the tabular presentations. In addition, generalized functions may facilitate secondary data analyses which were not conducted in the initial publications. Generalized variance functions may also provide more stable estimates of variance by diminishing the variability of the individual variance estimates themselves.

In this paper we will study some generalized models for the standard error of proportion estimates from the 1988 National Household Survey on Drug Abuse. A log-linear model based upon the concept of a design effect will be developed. The final model will be evaluated against the simple average design effect model.

2. General Models

Wolter (1985) summarizes many of the most commonly used models for generalized variance functions. In this paper, we will be concerned with standard error estimates of proportions estimated from an unequally weighted multistage sample survey. In developing our model we will use the concept of a design effect (DEFF) popularized by Kish (1965). The design effect is the ratio of the design-based variance to the variance that would have been obtained from a simple random sample (SRS) of the same size. For an estimate, α, we have

$$\text{DEFF}(\alpha) = \text{Var}(\alpha)/S^2(\alpha) \qquad (1)$$

where $\text{Var}(\alpha)$ is the designed-based variance of the estimate and $S^2(\alpha)$ is the SRS variance. For proportions, the SRS variance is

$$s^2(p) = p(1-p)/n \qquad (2)$$

where n is the sample size used to compute p. The design effect summarizes the effects (due to stratification, clustering and unequal weighting) on the variance of a complex sample design.

One simple model is to calculate an average design effect for specific analysis domains or types of outcome. The average design effect is then used as follows to approximate the standard error estimate

$$SE(p)_{appx} = \left[DEFF_{ave} * [p(1-p)]/n\right]^{1/2} \qquad (3)$$

where

p = the estimated proportion

n = the sample size used to calculate p

$DEFF_{ave}$ = the average DEFF for the domain of interest

$SE(p)_{appx}$ = the approximate standard error of p.

An alternative generalized variance model for proportions is obtained via direct regression modelling. The following relation for an individual design effect estimate serves as an inspiration for the theoretical model:

$$\text{DEFF} = \text{Var}(p) \;/\; [p(1-p)/n].$$

Taking the log (base 10) of both sides of the above relation leads to a parsimonious model for the variance or standard error of a proportion, which can be expressed in terms of the following log-linear model:

$$\log[SE(p)] = \beta_0 + \beta_1 \log(p) + \beta_2 \log(1-p) + \beta_3 \log(n), \qquad (4)$$

where

p = estimated proportion

SE(p) = design-based standard error estimate for the proportion,

n = number of respondents in the subgroup under investigation

β_0, β_1, β_2, β_3 = regression coefficients for the intercept, log(p), log(1-p), and log(n), respectively.

More complicated models which allow for separate slopes and intercepts by analysis domain and type of outcome were also explored.

3. The Data

The 1988 National Household Survey on Drug Abuse is a stratified multistage unequally weighted area household sample. In conjunction with the initial analysis of the data, several thousand estimated proportions and their design-based standard errors were calculated. These estimates consisted of 17 drug categories (e.g., marijuana, inhalants, cocaine, cigarettes, alcohol, etc.); 3 recency-of-use categories (ever used, used in past year, and used in past month); and 3 past year frequency-of-use categories for alcohol, cigarettes and marijuana (used at least once, used 12 or more times and used once a week or more). The proportions were estimated for domains defined by age, race and sex.

All of the estimates were available for use in this study. However, individual prevalence estimates for which the design effect was outside the range of 1-3,

or for which the relative standard error (RSE, the ratio of the standard error of the estimate vs. the estimate itself) was greater than 50%, were removed prior to the modelling process. Such extreme values for design effects and RSE's may reflect instability in the Taylor series estimate of the design-based variance (most often resulting from small sample sizes). These variance estimates were regarded as spurious and, therefore, were not included in the modelling process. With spurious estimates removed using the above rule, a total of 2,354 proportion estimates were used.

The design-based variance estimates were produced using RTI's survey data analysis software package SUDAAN (Shah, et. al 1989). SUDAAN uses a Taylor series linerization approach for variance estimation. For the NHSDA, with replacement selection of the primary sampling units was assumed.

4. Results

A table of age, race, and sex-specific average design effects appears in Table 1. A total of 36 cells are represented in the table, corresponding to marginals and cross-classifications of 4 age groups and 2 sexes for the total population (14 cells); 2 sex and 4 age group marginals within each of 3 race categories (18 cells); 3 race marginals; and 1 overall design effect for the total population.

The design effects of Table 1 can be used to calculate approximate standard error estimates using equation (3). Both domain-specific and overall design effect models for obtaining approximate standard error estimates were included in our evaluation, as discussed in Section 5.

Table 2 presents a summary of the log-linear regression model fitted to the NHSDA data. The fit explains most of the variation in the data, as witnessed by an R-squared value of 96.9%. Notice that the three slope coefficients associated with log(p), log(1-p), and log(n) are significantly different from 0.5 (or -0.5 for log(n)). This yields a significant improvement over a simple random sampling (SRS) model, where all three coefficients would be either 0.5 or -0.5. The SRS model only explains about 77% of the variation in standard error estimates.

We initially expected that separate domain effects may be needed to account for differential average cluster sizes across domains. Therefore, other more complicated versions of the log-linear model, such as those containing domain and recency-of-use effects, were investigated. While some of these effects were statistically significant, only modest gains in the amount of additional variation explained by them were obtained. Part of the effect of cluster size may be accounted for by the slope for log(n) being different from the SRS value of -0.5. As will be seen, the above specification of the log-linear model represents a concise, easy-to-use, yet relatively accurate, summary of the variation in standard error estimates. In addition, the simple regression model is generalizable to any domain and drug prevalence rate from the 1988 NHSDA not included in this exercise. Therefore, only the parsimonious specification of the log-linear model survived for further investigation.

Once the regression coefficients have been estimated (using lest squares methodology, for example), one can then substitute new values of p, (1-p), and n into the fitted model to obtain the predicted, or approximate, standard error of the prevalence rate via the following formula:

$$SE(p)_{appx} = \frac{10^{b_0} * p^{b_1} * (1-p)^{b_2}}{n^{-b_3}}, \quad (5)$$

where b_0, b_1, b_2, and b_3 are the fitted regression coefficients for the intercept, log(p), log(1-p) and log(n), respectively. Note that the model does not produce estimates which are symmetric about 0.50, or 50%, due to the specification of both p and (1-p), and the resulting difference in their regression coefficients.

To serve as a "control" model, we have calculated approximate standard errors under simple random sampling assumptions. Namely,

$$SE(p)_{appx} = \sqrt{\frac{p(1-p)}{n}}$$

where p, n, and $SE(p)_{appx}$ are defined as with model (3). Such a model is useful in evaluating improvements provided by other, more sophisticated, models which attempt to account for the complex sample design.

5. Evaluation

To reiterate, the following models were evaluated

- Average design effect
  - overall
  - domain-specific
- Log-linear model
- SRS control model

Two measures were used to evaluate the above models with respect to their predictive ability. The first of these, the model R-square statistic, is defined as

$$100 - \left[\frac{SS\ (Predicted - Actual)}{SS(Total)} * 100 \right]$$

where

Predicted	=	the predicted standard error from one of the models
Actual	=	the design-based standard error
SS(Predicted - Actual)	=	the sum of squared deviations between predicted and actual standard errors

SS(Total) = the sum of squared deviations between the actual (i.e., design-based) standard errors and their mean

R-square measures how well the predicted values correlate with the actual ones. Specifically, it represents the proportion of variation in standard error estimates "explained" by the model under consideration.

Another measure used to evaluate the models is the absolute relative difference (ARD), which is expressed as

$$\frac{|\text{Predicted} - \text{Actual}|}{\text{Actual}} * 100$$

where "predicted" and "actual" standard errors are defined as above. The mean ARD for the set of estimates quantifies the average distance (without regard for direction) between actual and predicted standard errors, expressed as a percentage of the actual standard errors. Smaller values for the mean ARD indicate a better fit. For each model described in the previous section, we have calculated mean ARD's for 1), the entire set of estimates, and 2), non-overlapping subsets of the estimates (specifically, age categories 12-17, 18-25, 26-34, and 35+).

Traditional "learning set vs. test set" methods were also employed to provide an objective evaluation of the models. In the first stage of such an evaluation, a set of prevalence estimates, known as the "learning set", is used to fit the models. In this study, the learning set estimates consisted of those prevalence rates and corresponding standard errors described in Section 3

At the second stage, the fitted models are compared by studying their behavior on a new set of estimates, called the "test set". Here, the test set consisted of another large set of 1988 NHSDA estimates. Estimates in the test set also related to drug prevalence and usage, but contained either new domains (e.g., population density) or new analysis variables (e.g., amount of drugs used per week, methods of cocaine administration) that were not included in the learning set described earlier. A total of 2,592 prevalence estimates were included in the test set.

The learning set models were evaluated on the basis of how well their predictions fit the actual, design-based standard errors for estimates in the test set. The measures of goodness of fit (i.e., model R-square, mean ARD, and age-specific mean ARD) were used to compare model-specific predicted vs actual (design-based) standard errors in the test set. Thus, for each of the models under consideration we have 12 measures of its performance: its R-square for both the learning and test set, as well as its total sample and age-specific mean ARD for both the learning and test sets. In this way, we could evaluate the predictive ability of our original learning set models in situations similar to those that may be encountered by future users of a database.

As a final note on goodness-of-fit measures, the domain-specific ARD's were limited to categories of age, primarily since age was the only domain which appeared consistently in both learning and test sets, thus yielding the only possible domain-specific comparisons between them.

The results of the evaluation are shown in Table 3. As expected, both the learning and test set R-square's were larger for the domain-specific average design effect (DEFF) model than for the overall average design effect model. Specifically, domain-specific DEFF R-square values were approximately 94% and 90% for learning and test sets, while the overall DEFF R-square's dropped to 90% and 86% for the learning and test sets, respectively.

Just as the R-square values dropped in the test set for both average DEFF models, the mean absolute relative differences (ARD's) increased. As anticipated, total sample and age-specific mean ARD's were universally smaller for the domain-specific DEFF model, in both learning and test sets. Overall, the mean ARD's for the domain-specific DEFF model ranged from 11% (learning set values) to 26% (maximum test set values); such values ranged from 15% to a maximum of 34% using the overall DEFF model.

Table 3 also contains the goodness of fit measures for the log-linear model. Note that the R-square values for the log-linear model, 97% and 93% (learning and test datasets) are slightly larger than the values for the domain-specific average DEFF values (94% and 90%, respectively). Also, the mean absolute relative differences (ARD's) for the log-linear model are close to the values for the domain-specific average DEFF model. The mean ARD's for the learning data set are slightly smaller for the domain-specific average DEFF model, while for the test dataset the log-linear model mean ARD's are slightly lower.

Results for the control model, using simple random sample assumptions to obtain predicted standard errors for both learning and test sets, are also presented in Table 3. This model is provided to serve as a reference, against which we can evaluate the improvement provided by other, more sophisticated, models. R-square values for the control model are 77% and 84% for the learning and test sets, respectively. ARD values range from 19% to 22% in the learning set, and from 21% to 24% in the test set. Recall that R-square values were greater than 90% for the log-linear model, and ARD's were one-third smaller than the SRS mean ARD's.

6. Conclusions

We conclude that, for these data, the domain-specific average DEFF model (equation 3) and the simple log-linear model (equation 5) both provide adequate generalized standard error models. We were surprised that the log-linear model including only effects for log(p), log(1-p), and log(n) preformed so well. We expected domain effects would be required in the model to account for differential average cluster sizes by domain for this multistage sample design. It appears that the slope for log(n) in the model being different from the SRS value of -0.5 accounts for most of the cluster size effect.

To further see this point, a model predicted design effect equation can be constructed by dividing the predicted variance by the SRS variance, p(1-p)/n. This yields

$$DEFF = 1.2753\ p^{.062}\ (1-p)^{.109}\ n^{.0668}.$$

Figure 1 plots this function for proportions of 0.05 and 0.50 by sample size. As the sample size (and, hence, the average cluster size) increases, the predicted design effect increases. Also, the design effect is larger for proportions near one half and smaller for proportions near zero (or one).

7. References

Shah, Baba V., et. al (1989). SUDAAN: Procedures for Descriptive Statistics, Research Triangle Institute, Research Triangle Park, North Carolina.

Wolter, Kirk M. (1985). Introduction to Variance Estimation, Springe-Verlog, New York, New York.

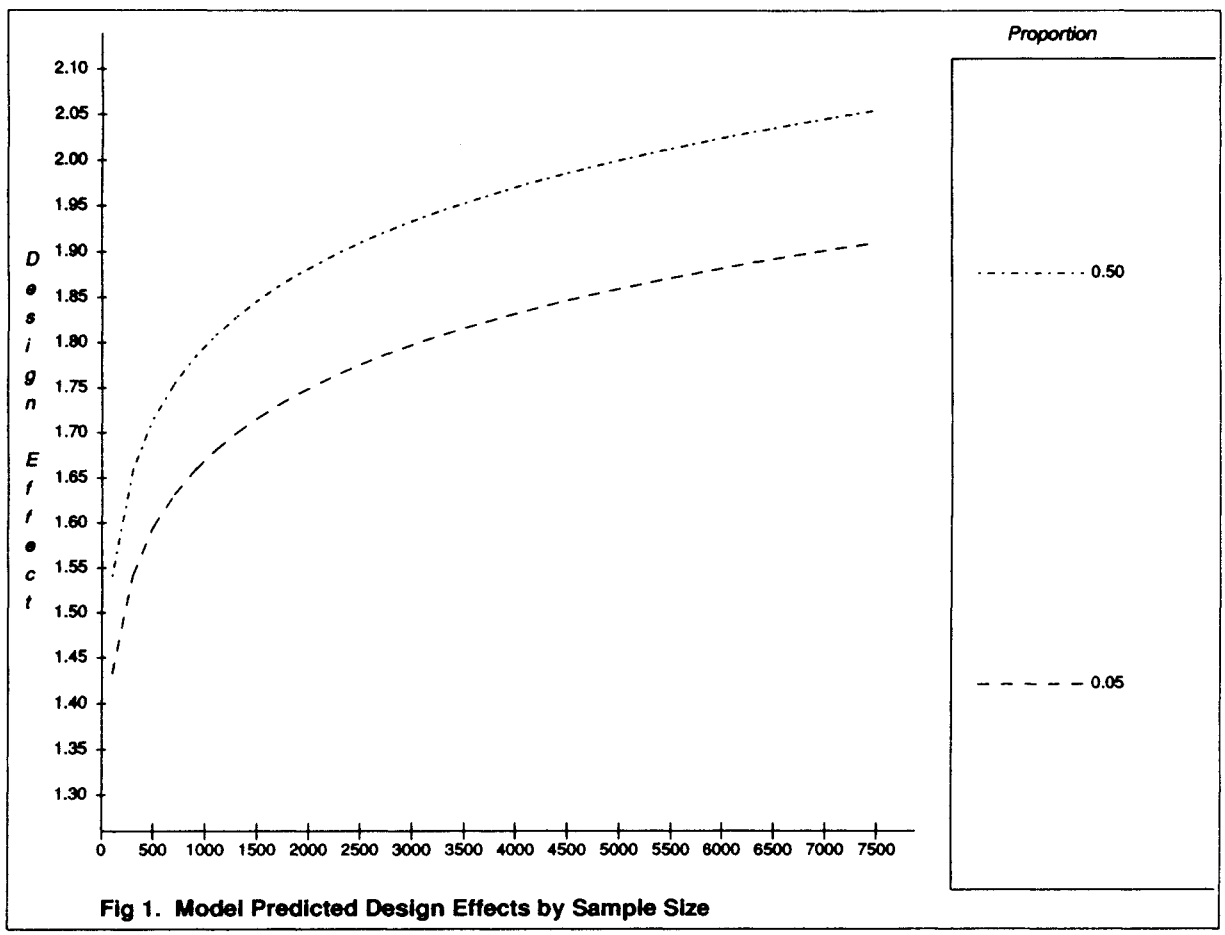

Fig 1. Model Predicted Design Effects by Sample Size

Table 1. Average Design Effects by Sex, Age Group, and Race

Sex	Age/Race Marginals		Age/Race Cross-classifications	
Males	12-17	1.44	12-17	
	18-25	1.68	Whites	1.51
	26-34	1.65	Blacks	1.32
	35+	1.52	Hispanics	1.47
	Whites	1.64	All Races	1.87
	Blacks	1.69		
	Hispanics	1.69	18-25	
			Whites	1.43
	All Males	1.98	Blacks	1.50
			Hispanics	1.61
Females	12-17	1.83		
	18-25	1.80	All Races	1.91
	26-34	1.64		
	35+	1.89	26-34	
			Whites	1.54
	Whites	1.75	Blacks	1.64
	Blacks	1.66	Hispanics	1.88
	Hispanics	2.07		
			All Races	1.84
	All Females	2.07		
			35+	
Total	12-17	1.87	Whites	1.33
	18-25	1.91	Blacks	1.28
	26-34	1.84	Hispanics	1.68
	35+	1.67		
			All Races	1.67
	Whites	1.82		
	Blacks	1.83		
	Hispanics	1.96		
Overall		2.06		

Source: National Institute on Drug Abuse, 1988 National Houshold Survey on Drug Abuse.

Table 2. Simple Log-linear Regression Model Results

Variable	Beta	Standard Error	95% C.I. Lower	95% C.I. Upper	R-Squared (%)	P-Value
Intercept	0.0528	0.0116	0.0301	0.0755	96.9	0.0001
Log(p)	0.5310	0.0030	0.5251	0.5369		
Log(q)	0.5545	0.0091	0.5367	0.5723		
Log(n)	-0.4666	0.0038	-0.4740	-0.4592		

Table 3. Model Evaluation Results

Model	R-Square		Total Sample ARD		Age-Specific* ARD							
	Learning	Test	Learning	Test	Learning				Test			
					1	2	3	4	1	2	3	4
Average design effects												
• Overall	90	86	16	24	16	16	16	18	21	20	24	35
• Domain-specific	94	90	11	21	10	11	11	10	17	18	20	26
Log-linear model	97	93	11	17	11	12	11	12	13	15	16	22
SRS control model	77	84	22	24	21	22	21	20	21	24	23	23

* 1 = 12-17 yrs.
2 = 18-25 yrs.
3 = 26-34 yrs.
4 = 35+

Source: 1988 National Household Survey on Drug Abuse, National Institute on Drug Abuse.