GENERALIZED VARIANCE FUNCTION METHODOLOGY FOR ACNIELSEN'S HOMESCAN HOUSEHOLD PANEL SURVEY

Don Jang, Joseph K. Garrett, Mathematica Policy Research, Inc.
Frank W. Piotrowski, William B. Owens, ACNielsen

Key Words: Generalized variance function, GVF diagnostic statistics, panel survey, variance estimation

ACNielsen's Household Panel survey consists of a sample of 40,000 households selected throughout the contiguous U.S. For selected households, ACNielsen provides in-home, electronic data capture scanners so panelists can scan all bar-coded purchases made by household members. Once a week, panelists transmit the captured data over telephone lines to ACNielsen processing centers. These survey data provide information on consumer demographics, items purchased, frequency and quantity of purchases, costs of purchases, and locations of purchases, among other things so that marketers can understand the dynamics associated with consumers' buying behavior and shopping patterns.

This household panel survey was designed to provide market-level estimates for 16 pre-defined major markets as well as national estimates. To assess the reliability of survey estimates, measures of precision (such as standard errors or variances of survey estimates) need to be presented. However, it is complicated to compute variance estimates from this household panel survey data for several reasons. First, the final weights for this panel are constructed through a complicated, multi-dimensional raking procedure to adjust to estimated population counts by demographic level and by household level. This weighting adjustment is periodically implemented to adjust updated population counts and updated panels. Second, the household panel service produces literally thousands of estimates on an ongoing basis, making implementation of direct methods of variance estimation a cumbersome and expensive process. And finally, most client reports have customized tables of estimates, further complicating standardized variance estimation.

Consequently, it has been desirable to have a procedure capable of producing estimates of precision for many estimates in a timely, inexpensive manner. Additionally, the procedure should be user-friendly so that nonstatisticians and analysts can compute estimates of precision as needed and with minimal training.

To accomplish these objectives, we developed generalized variance functions (GVFs) for four statistics presented below.

1. Development of GVFs

A generalized variance function (GVF) is a mathematical model that describes the relationship between a population value to be estimated (such as a population total) and its estimator's variance. Detailed discussions for GVFs are found in Chapter 5 of Wolter 1985.

For the ACNielsen's Household Panel Survey, we developed GVF models for these four types of key statistics:

- Total: projected total purchase level of a product item
- Share: projected share level (in percentage) of a product item among the product category including all items
- Buying rate: the ratio of projected purchases to projected buyers
- Penetration rate: percentage of households purchasing the item

We restrict our attention to six product categories: carbonated beverages, cereal, dentifrice, deodorant, shampoo, and snacks. Four geographic areas are considered for this study: contiguous U.S., Boston, Houston, and Seattle. All data are for the second quarter of 1998.

Overview

We developed generalized variance functions for *quarterly estimates* of the four key statistics. Given a form of GVF model corresponding to the statistic, separate model fitting was executed for four geographic areas by six product categories. In addition, we also produced a combined GVF derived from all product items across all six product categories for each of the four

geographic areas. These GVFs can then be used to approximate standard errors of statistics for other product categories not used in this GVF development, if possible. Furthermore, all three markets and six product categories were combined to get a GVF which can be used for other product categories and other market areas.

Steps for GVF Developments

GVF modeling consists of two steps. First, estimates of population values (such as total purchasing levels, share levels, buying rates, and penetration rates) and their estimated variances were calculated directly for all variables in the data set. Next, we chose two independent sets of variables for each GVF development. To do so, we ordered all variables by their magnitudes and selected two sets of variables alternately so that these variables represent the range of all variables. The first set of variables was used to develop the initial GVF models. The second set was used to evaluate the quality of the generalized standard errors in comparison to the directly estimated standard errors. Details of the direct variance calculation for this study were presented in Jang and Garrett 1999.

The next step was to use the estimated values of key statistics and their associated variances to create models that allow users to predict the standard error for a statistic that they have estimated. Once developed and validated, the model can be used to estimate variances for other variables for which direct estimates of variance have not been computed.

GVF functional forms are different for the types of statistics we considered.

Total Purchasing Levels

Let $\hat{Y}$ denote an estimator of the population total purchase level Y. GVF models are usually created for the relative variance of an estimated total $\hat{Y}$, or $RelVar(\hat{Y}) = Var(\hat{Y})/Y^2$, where $Var(\hat{Y})$ is the variance of $\hat{Y}$. Modeling typically begins by assuming that the relative variance of the estimated total $\hat{Y}$ decreases as the total Y increases. Empirical investigations supported hypothesis that the log-transformation of both the relative variance and the total would give a better relationship:

$$\text{Log}\{RelVar(\hat{Y})\} = \beta_0 + \beta_1 \text{Log}(Y) \qquad (1)$$

A usual least squares fit was used to obtain model parameter estimates.

The relative variance of an estimated total $\hat{Y}$ can be predicted by evaluating the GVF model at $\hat{Y}$ and at $\hat{\beta}_0$ and $\hat{\beta}_1$ which are the estimates of the model parameters β_0 and β_1 from GVF model (1) and then transposing the model to get:

$$se(\hat{Y}) = e^{\frac{\hat{\beta}_0}{2}} \hat{Y}^{1+\frac{\hat{\beta}_1}{2}} \qquad (2)$$

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

Share Levels

Share level P is the percentage of projected total purchasing level (X) of the product item among all projected purchasing level (Y) of the product category, i.e., $P = 100Y^{-1}X$. It may be plausible to assume that an estimated share level $\hat{P}$ is independent of the projected total purchasing level ($\hat{X}$) of the product category ($\hat{Y}$), where $\hat{P} = 100\hat{Y}^{-1}\hat{X}$. Consequently, the resulting relative variance of $\hat{P}$ can be simplified as:

$$RelVar(\hat{P}) = RelVar(\hat{X}) - RelVar(\hat{Y}) \qquad (3)$$

Since relative variances of $\hat{X}$ and $\hat{Y}$ can be obtained from the same GVF model in (1), the relative variance of the share level $\hat{P}$ can thus be calculated.

Once the models for total purchasing levels are adapted, the same models can be used to predict standard errors for share levels. Using the parameter estimates from the models in (1), the standard error for share levels can be predicted with this formula:

$$se(\hat{P}) = \hat{P}e^{\hat{\beta}_0/2}\hat{Y}^{\hat{\beta}_1/2}\{(\hat{P}/100)^{\hat{\beta}_1} - 1\}^{1/2} \qquad (4)$$

where $se(\hat{P})$ is the GVF-predicted standard error for a specific estimated share level $\hat{P}$ and $\hat{Y}$ is the projected total purchasing level of the product category as the base of the share level.

Penetration Rates

Penetration rate P is the percentage of households purchasing the product item. Like the GVF approach for share levels, GVFs for penetration rates can be derived from those for projected buyers. The same form of the GVF model as in (1) was adapted to fit models for projected number of buyers with strong empirical evidence of that relationship.

Like the case of share level, we can set the assumption of the independence between an estimated penetration rate $\hat{P}$ and projected total U.S. household level buyers $\hat{Y}$. All remaining GVF procedures are then equivalent as in estimates for share levels. That is, the relative variance of an estimated penetration percentage $\hat{P}$ can be expressed as:

$$RelVar(\hat{P}) = RelVar(\hat{X}) - RelVar(\hat{Y}) \qquad (4)$$

where $\hat{X}$ is projected total number of buyers for a specific item and $\hat{Y}$ is the projected total population household size in U.S. Consequently, using the parameter estimates from the models in (2), the standard error for penetration rates can be predicted with this formula:

$$se(\hat{P}) = \hat{P} e^{\hat{\beta}_0/2} \hat{Y}^{\hat{\beta}_1/2} \{(\hat{P}/100)^{\hat{\beta}_1} - 1\}^{1/2} \qquad (5)$$

where $se(\hat{P})$ is the predicted standard error for a specific estimated penetration percentage $\hat{P}$ and $\hat{Y}$ is the estimated total U.S. household size as the base of the percentage.

Buying Rates

Buying rate is a ratio statistic in which the numerator and denominator are both random variables and are measured with different units. Unlike the two percentage statistics presented above, the independence assumption of numerator and denominator cannot be made. We believe there is a somewhat positive relationship between them. Actual empirical investigation using the data supplied supports this. This means the approach used for previous two percentage values is not appropriate for buying rates. We attempted to derive appropriate models using empirical plots of key statistics versus variance estimates, or transformations of those.

Fortunately, we were able to identify reasonable models for this ratio. Empirical scatter plot results showed that the relative variance of buying rate $\hat{R}$ has a certain relationship with the product of buying rate and penetration rate. Since the log-transformation can make the product term linear, we used a log-transformation. Specifically, we set the model:

$$\text{Log}\{SE(\hat{B})\} = \beta_0 + \beta_1 \text{Log}(B) + \beta_2 \text{Log}(P). \qquad (6)$$

With model parameter estimates from the model in (6), the GVF-based standard errors of buying rates can be obtained:

$$se(\hat{B}) = e^{\hat{\beta}_0} \hat{B}^{\hat{\beta}_1} \hat{P}^{\hat{\beta}_2}. \qquad (7)$$

where $se(\hat{B})$ is the predicted standard error for an estimated buying rate $\hat{B}$ and $\hat{P}$ is an estimated penetration rate for a specific product item.

2. Evaluation of GVF Models

To assess whether the model provides adequate predictions of variance estimates, we evaluated GVF-driven standard errors using the diagnostic called the absolute value of the relative error *AREL-ERR:*

$$AREL\text{-}ERR = \left| \frac{SE_{\text{Act}} - SE_{\text{Pred}}}{SE_{\text{Act}}} \right| \qquad (8)$$

where SE_{Act} is the directly calculated standard error and SE_{Pred} is the GVF-predicted standard error. *AREL-ERR* measures the relative loss of accuracy due to using GVF approximations.

The average *AREL-ERR* quantifies the average distance between actual versus predicted standard error, which we express as a relative difference of the actual standard error. Small values for the average *AREL-ERR* indicate that the corresponding GVF can be reliably used. We produce these measures for a set of variables used for fitting the model (MODEL) and a set of variables which were initially not included in fitting the model (TEST) to see how reliable the resultant model would be. There is little difference between values from MODEL and those from TEST for all four statistics.

3. GVF Products

After evaluating and confirming models, GVF-based standard errors can be obtained.

Total Purchasing Levels

The following steps allow to approximate the standard error of an estimated total purchase:

- Obtain the estimated total purchasing levels of an item for a specific market (or national level)
- Determine whether to use a product-specific GVF or the combined GVF: we recommend to use product-specific GVFs for total purchasing levels
- Get estimates of parameters for the chosen GVF

- Compute the GVF-driven standard error using the formula (2)

For example, the U.S.-level projected total purchasing level of "total carbonated beverages and FD/FJ" is $\hat{Y}=2.74\times10^{14}$. From GVF model fitting, the corresponding parameter estimates are $\hat{\beta}_0=12.732$ and $\hat{\beta}_1=-0.64$. Thus, a GVF-driven standard error of the U.S.-level projected total purchase level is:

$$se(\hat{Y}) = e^{\hat{\beta}_0/2}\hat{Y}^{1+\hat{\beta}_1/2} = e^{12.73/2}\times(2.74\times10^{14})^{1-0.64/2} = 3.52\times10^{12}.$$

With a directly calculated standard error of 3.21×10^{12}, the relative error of this predicted value would be 0.1.

Share Levels

The following steps allow to approximate the standard error of an estimated share level:

- Obtain the estimated total purchasing level for all items in the product category and share percentage of an item for a specific market (or national level)
- Determine whether to use a product-specific GVF or the combined GVF (like total purchasing levels, we recommend to use product-specific GVFs)
- Get estimates of parameters for the chosen GVF
- Compute the GVF-driven standard error using the formula (5)

For example, the U.S. level share percentage estimate of carbonated beverages among the carbonated beverage category is 66.2 percent. Then, we need to know the U.S. level total purchasing level for the carbonated beverage product category to which the item belongs. In the second quarter of 1998, the value was 2.74×10^{14}. With parameter estimates of $\hat{\beta}_0=12.73$ and $\hat{\beta}_1=-0.64$, an approximate standard error of an estimate share level is:

$$\begin{aligned} se(\hat{P}) &= \hat{P}e^{\hat{\beta}_0/2}\hat{Y}^{\hat{\beta}_1/2}\{(\hat{P}/100)^{\hat{\beta}_1}-1\}^{1/2} \\ &= 66.2\times e^{12.73/2}\times(2.74\times10^{14})^{-0.64/2}\times(0.662^{0.64}-1)^{0.5} \\ &= 0.47 \end{aligned}$$

In fact, a directly calculated standard error is 0.37 with the relative error of -0.27.

Penetration Rates

The following steps allow to approximate the standard error of an estimated penetration rate:

- Obtain the projected total buyers in a market or U.S. level depending on whether a penetration rate of interest is a market level or U.S. level
- Calculate penetration rate of an product item within a market area (or U.S.)
- Determine whether to use a product-specific GVF or the combined GVF (combined GVFs would be preferable)
- Get estimates of parameters for the chosen GVF
- Compute the GVF-driven standard error using the formula (5)

Suppose we are interested in estimating penetration rate of carbonated beverages among all U.S. households. A calculated penetration rate during the second quarter of 1998 was 91.3%. Then, we also need to know the total number of U.S. household to get GVF-driven standard errors. In the second quarter of 1998, the estimated household size was 101,041,273. The combined U.S. level GVF can be used. With parameter estimates of $\hat{\beta}_0=8.49$ and $\hat{\beta}_1=-0.96$, an approximate standard error of an estimated penetration rate is:

$$\begin{aligned} se(\hat{P}) &= \hat{P}e^{\hat{\beta}_0/2}\hat{Y}^{\hat{\beta}_1/2}\{(\hat{P}/100)^{\hat{\beta}_1}-1\}^{1/2} \\ &= 91.3\times e^{8.49/2}\times101,041,273^{-0.96/2}(0.913^{0.96}-1)^{0.5} \\ &= 0.28 \end{aligned}$$

With a directly calculated standard error of 0.27, the relative error of the GVF-driven one is -0.02.

Buying Rates

The following steps allow to approximate the standard error of an estimated buying rate:

- Obtain the estimated buying rate and penetration rate of an item for a specific market (or national level)

- Determine whether to use a product-specific GVF or the combined GVF
- Get estimates of parameters for the chosen GVF
- Compute the GVF-driven standard error using the formula (7)

For example, the U.S. level buying rate estimate of all carbonated beverages is 1,966,654. To get GVF-driven standard error, we need to have penetration rate estimate for these carbonated beverages also. The corresponding penetration rate is 91.3%. With parameter estimates of $\hat{\beta}_0 = -3.42$, $\hat{\beta}_1 = 1.08$, and $\hat{\beta}_2 = -0.34$, an approximate standard error of an estimate buying rate is:

$$se(\hat{B}) = e^{\hat{\beta}_0} \hat{B}^{\hat{\beta}_1} \hat{P}^{\hat{\beta}_2} = e^{-3.42} \times 1,966,654^{1.08} \times 91.3^{-0.34} = 41,456.$$

With a directly calculated standard error of 22,322, the relative error of the GVF-driven one is -0.86.

4. Conclusions

All final models were also evaluated with a diagnostic presented in this paper. For each of the four statistics we considered thus far, we summarize the results below.

Total Purchasing and Share Levels. Due to the difference of scales according to product categories, it turned out to be difficult to combine GVFs across different scales. At this moment, we may suggest to use product-specific GVFs to predict standard errors of these two statistics for product categories considered. One suggestion we can make for the future project is to group the whole product categories into several super product categories with respect to their scales. With this approach, several product-combined GVFs can be produced and ultimately cover all product categories.

Penetration Rate. Variance or standard error calculation of a penetration rate can be simplified by using U.S. or market level combined GVFs. Specifically, for any market-level penetration rate, we would use the formula (5) with $\hat{\beta}_0 = 6.76$ and $\hat{\beta}_1 = -0.93$. Similarly, for any U.S. level penetration rate, we can use the same formula with $\hat{\beta}_0 = 8.49$ and $\hat{\beta}_1 = -0.96$.

Buying Rates. GVF-driven standard errors for buying rates seem to be less accurate in a sense that they have larger deviation from the actual standard errors than other statistic. The final models we fitted is quite empirical and intuitive. However, the proposed model is still attractive to be used because it is simple to implement and the deviation from the actual values still seems to be acceptable. Once users are willing to use the model, we suggest to use the combined models rather than product-specific models.

Other Issues. Periodically, the models would need to be updated to ensure they continue to capture the inherent sampling variability associated with the household panel estimates. An extension of this procedure can be made to compute variance estimates at levels other than nationally and three major market areas, for time periods other than quarters, and for estimates other than the four key statistics.

REFERENCES

Jang, Don and Joseph K. Garrett (1999). "Variance Estimation Methodology for ACNielsen's Homescan Panel Survey." Washington, D.C.: Mathematica Policy Research.

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