GENERALIZED VARIANCE FUNCTIONS FOR THE CURRENT EMPLOYMENT STATISTICS SURVEY

Steve Woodruff, Bureau of Labor Statistics
Bureau of Labor Statistics, Rm 4985-06, 2 Massachusetts Ave. N.E., Washington D.C. 20212

1. INTRODUCTION

The Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) Survey gathers data monthly from over 380,000 nonagricultural business establishments allocated across all States, for the purpose of estimating total employment, employment of women and production workers, hours, and earnings. Estimates are made for over 1,500 industry cells, complementing the demographic detail provided by estimates of employment from the Current Population Survey (CPS). Monthly estimates of level and over-the-month change in employment are of primary importance to the users of these data. In addition to national estimates of level and change for these 1500 cells, state estimates for many of these cells are required. Unfortunately, many estimation cells within individual states, suffer sample sizes that are extremely small. It is not necessarily the case that variances of such cell estimates are always going to be large; there is a complex relationship between cell variances of employment estimates and cell characteristics other than sample size. This paper documents the derivation of these relationships and their application to variance estimation using generalized variance functions (GVFs).

The estimate of over-the-month change for a basic estimating cell is the quotient of total employees for the current month in the matched sample (all sample establishments that reported employment for both current and previous month) over total employees for the previous month in the matched sample. This quotient is called the link (between current and previous month). The estimate of level for the current month is the product of this link and the estimate of level for the previous month. This estimate of level is called the link relative estimator. It is closely related to both the ratio and the regression estimator.

Royall and Cumberland (1978,1981), and Royall and Eberhardt (1975) looked at the general problem of estimating the variances of ratio and regression estimators. Their findings lead to some of the estimators tested here. The specific problem of estimation of CES variances has been studied by West (1984), Royall (1981), and Madow & Madow (1978). The estimators considered here include variations on their suggestions, with the emphasis on computational simplicity.

The derivation of the generalized variance functions follow the motivation and goals outlined in Valliant (1992): variance estimates should be computationally simple, be approximately unbiased, yield reasonable confidence interval coverage, and exhibit greater stability than the point estimates of variance. If such estimators can be derived, they will be used to establish publishability in place of the present rule that requires either a cell sample size of at least 15 or that the cell sample employment be 50% or more of the total cell employment.

A universe data base was created for this study from the California and Michigan Unemployment Insurance (UI) reports covering the 18-month period of Jan 90 to Jun 91. For the California data, within each of 352 three digit SICs, sampling and estimation were replicated 50 times following as nearly as possible the actual CES sampling and estimation methodology (the link relative estimator) . These 50 replicate estimates were then used to estimate the relative variance of the link relative estimator in each 3-digit SIC cell over a nine month period (April 1990 to Dec 1990). This gave us 352x9=3168 relative variance estimates by SIC cell and month to be used to fit GVFs of cell characteristics and time (months from benchmark month).

The generalized variance function suggested for implementation is:

$$RV = \mathrm{Exp}(\mathrm{Log}(g) - \mathrm{Log}(b) + (2/3)\mathrm{Log}(H) + (2/3))$$
$$= (1.93)(g/b)H^{(2/3)}$$

where RV is the relative variance of the link relative estimator of total employment (see 2.3) in an estimation cell, b is the total sample employment (total number of employees in the sample units), H is the number of months between estimation month and benchmark month (See equation 2.3), and g is

the cell's finite population correction factor (g=1-(b/LR) where LR is the estimate of cell employment.

2. THE LINK RELATIVE ESTIMATOR.

The over-the-month link is the ratio of the total matched sample employment for the current month over the total matched sample employment for the previous month, where the "matched sample" consists of those units which reported employment for both months. The over-the-month link is an estimate of employment change between adjacent months.

For a given pair of adjacent months j and j-1, let a_j be the total employment in the matched sample for month j. Let b_j be the total employment in the matched sample for the previous month, j-1. The over-the-month link is $\hat{\beta}_j = a_j/b_j$. Under the model that describes an establishment's total employment in the CES survey we have:

$$a_j = \beta_j b_j + \varepsilon_j \qquad (2.1)$$

where $E(\varepsilon_j) = 0$, β_j is an unknown constant (estimated with $\hat{\beta}_j$). $Var(\varepsilon_j|b_j)=Kb_j$, where K is an unknown constant, and a_j & b_j are given as follows:

$$a_j = \sum_{k \varepsilon s_j} y_{kj} \quad \& \quad b_j = \sum_{k \varepsilon s_j} x_{kj}$$

s_j is the matched sample for months j and j-1, x_{kj} is the total number of employees (all employment) in the kth sample establishment for month j-1, and y_{kj} is all employment in the kth sample establishment for the current (reference) month j.

This model implies that the conditional variance of the link given b_j is:

$$V([\hat{\beta}_j]|b_j) = K/b_j \qquad (2.2)$$

The CES estimate for total cell employment in month H is called the link relative estimator (LR_H) and is given as:

$$LR_H = (BM)\prod_{j=1}^{H}\hat{\beta}_j \qquad (2.3)$$

where (BM) is actual total employment for the base period (the benchmark month) and the $\{\hat{\beta}_j\}$ are the month-to-month estimates of change between months j and j-1. The benchmark month is either the most recent March or the March before that. The link relative estimator is the "Best Linear Unbiased Estimator" under (2.1) , (2.2) , and the feature of CES data response called hierarchical data flow. Hierarchical data flow occurs when the set of respondents for month j is contained in the set for month j-1, for each month back to the benchmark month. The shuttle schedule used in the CES survey encourages hierarchical data flow.

An identical estimation process is carried out in most of the 1500 estimation cells. A subscript denoting cell was purposely omitted to simplify notation.

3. THE RELATIVE VARIANCE OF THE LINK RELATIVE ESTIMATOR

Consider relationships between the conditional variance of the month-to-month links:

$$V(\hat{\beta}_j|b_j)=g_jK/b_j, \qquad (3.1)$$

where K is a constant and g_j is a finite population correction factor for month j. Recall from the previous section that, a_j is the matched sample employment at time j and b_j is the matched sample employment for time j-1. The matched sample consists of those units which were in the sample at both times j and j-1.

The above expression for variance is suggested by Cochran (page 153) and in Section 2, with the addition of a finite population correction factor. The variance of a ratio can be approximated as $(g_j/b_j)K$ where K is a constant independent of the sample outcome.

Gibrat (1930) suggests that when an estimation cell is selected at random (equal probability) from among all the CES estimation cells and its employment is observed then this employment value is approximately lognormally distributed. This property seems to hold for matched sample employment too. Taking logs transforms the empirical densities of both matched sample employment and estimated employment (as well as, estimated relative variance across cells) from highly skewed to roughly bell shaped.

The relative variance of a random variable can often be approximated by the variance of the Log of that random variable (when these items are

defined). This property holds for the link relative estimator (LR_H). Thus the variance of the Log of the link relative estimator is approximately equal to the relative variance of the link relative estimator.

The $\{\hat{\beta}_j\}$ are the sample links between adjacent months, BM is the cell benchmark employment, and

$$LR_H = (BM)\prod_{j=1}^{H}\hat{\beta}_j ,$$

Then :

$$RV(LR_H) \doteq V(Log(LR_H)) =$$

$$V(Log(BM) + \sum Log(\hat{\beta}_j)),$$

The relative variance of the link relative estimator can be expressed as the sum of variances of the Logs of these links and covariances between them (Log(BM) is constant).

$$RV(LR) = V(Log(BM) + \sum Log(\hat{\beta}_j))$$

$$= \sum_{j=1}^{H} V(Log(\hat{\beta}_j)) + \sum_{i\neq k}^{H} Cov(Log(\hat{\beta}_i), Log(\hat{\beta}_k)) \quad (3.2)$$

The autocorrelations between the $\{\hat{\beta}_j\}$ approximate the autocorrelations between the $\{Log(\hat{\beta}_j)\}$ since for $|\hat{\beta}_j - 1|$ small, $Log(\hat{\beta}_j) \doteq \hat{\beta}_j - 1$. The autocorrelations between the $\{\hat{\beta}_j\}$ are a decreasing nonnegative function, r, of the number of months separating the two sample links. For example, the correlation between $\hat{\beta}_i$ & $\hat{\beta}_k$ (or between $Log(\hat{\beta}_i)$ & $Log(\hat{\beta}_k)$) is:

$$\rho_{\hat{\beta}_i, \hat{\beta}_k} = r(|i-k|).$$

(3.2) becomes:

$$RV(LR) = \sum_{j=1}^{H} V(Log(\hat{\beta}_j)) +$$

$$\sum_{i\neq k}^{H} Cov(Log(\hat{\beta}_i), Log(\hat{\beta}_k)) =$$

$$\sum_{j=1}^{H} V(Log(\hat{\beta}_j)) +$$

$$\sum_{i\neq k}^{H} r(|i-k|)\sqrt{V(Log(\hat{\beta}_i))V(Log(\beta_k))} \quad (3.3)$$

Since the month-to-month links, $\{\hat{\beta}_j\}$, are ratios of sample employments (lognormally distributed), the links are also lognormal, $Log(\hat{\beta}_j)$ ~ $N(0, \sigma_j^2)$, where σ_j^2 is unknown and since the month-to-month link vary about one, their logs vary about 0. Then, by Lognormality, Johnson and Kotz (1970) pp 115, the variance of a sample link can be written as: $V(\hat{\beta}_j) = e^{\sigma_j^2}(e^{\sigma_j^2} - 1)$ and under historical experience with the CES data (also Cochran), $V(\hat{\beta}_j) = (Kg_j)/b_j$, where K is a constant, g_j =(1-(b_j/LR_j)) , is a finite population correction factor, b_j is the sample employment in the denominator of the j^{th} sample link.

If these two expressions for $V(\hat{\beta}_j)$ are equated and solved for $e^{\sigma_j^2}$ in terms of b_j and g_j , the result is: $e^{\sigma_j^2} = (1/2) + \sqrt{(1/4) + (Kg_j/b_j)}$, or:

$$\sigma_j^2 = Log[(1/2) + \sqrt{(1/4) + (Kg_j/b_j)}],$$

For a given estimation cell, both g_j and b_j remain relatively constant over time (j), dropping their subscripts acknowledges this fact and σ_j^2 can be written as:

$$\sigma_j^2 = Log[(1/2) + \sqrt{(1/4) + (Kg/b)}] \quad (3.4)$$

Thus (3.3) can be written as:

$$RV(LR)= \sum_{j=1}^{H}\sigma_j^2 + \sum_{i\neq k}^{H} r(|i-k|)\sigma_i\sigma_k \ ,$$

and substituting (3.4), RV(LR) =

$$\sum_{j=1}^{H} Log[(1/2)+\sqrt{(1/4)+(Kg/b)}] +$$

$$\sum_{i\neq k}^{H} r(|i-k|)Log[(1/2)+\sqrt{(1/4)+(Kg/b)}]$$

$$= Log[(1/2)+\sqrt{(1/4)+(Kg/b)}]\times \left(H+\sum_{i\neq k}^{H} r(|i-k|)\right) \qquad (3.5)$$

The relative variance of the link relative estimator are themselves very skewed; most cells have relative variances near zero and a few values are scattered about well above zero. This suggests that Log(RV(LR)) be considered. In fact, the log transformation results in the {Log(RV(LR))} being roughly bell shaped across estimation cells. Next it is shown that this additional transformation of the data simplifies the generalized variance function.

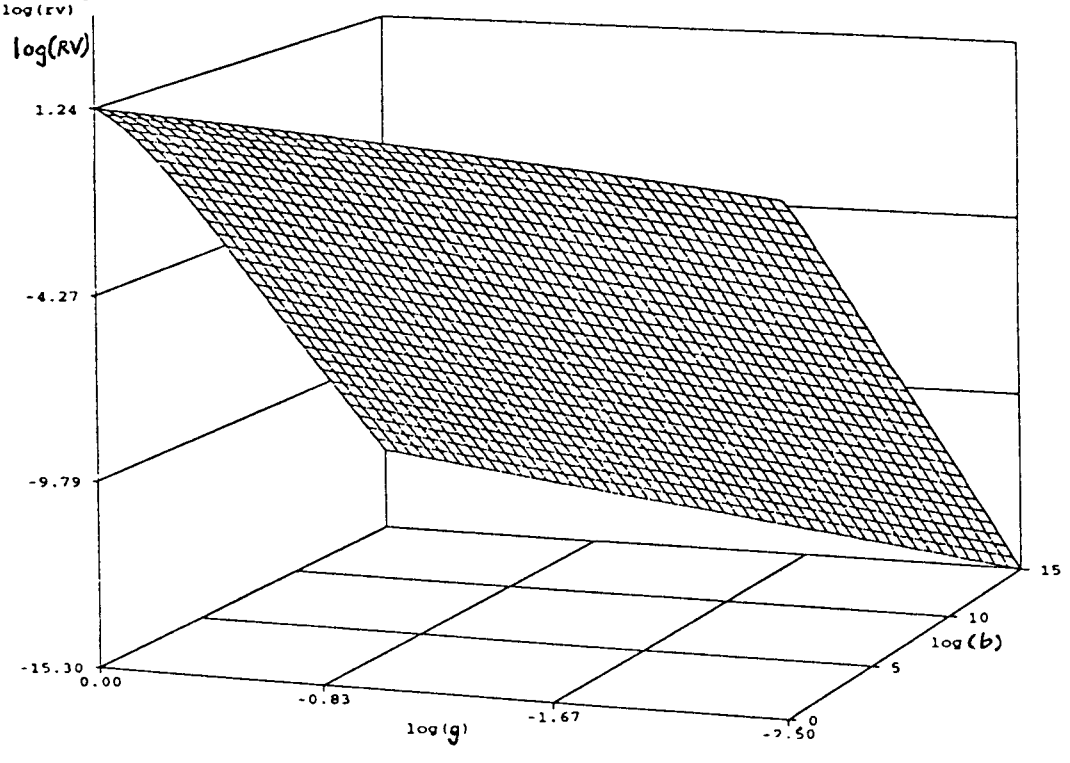
Taking Logs of both sides of 3.5 we have:

Log(RV(LR)) =

$$Log(H+\sum_{i\neq k}^{H} r(|i-k|)) + Log[Log[(1/2)+\sqrt{(1/4)+(Kg/b)}]] \qquad (3.6)$$

The Log of the relative variance is a function of H plus this iterated logarithm function of g and b. Although this iterated logarithm is a nonlinear function of g and b, if it is written as a function of Log(g) and Log(b) by substituting exp(Log(g)) for g and exp(Log(b)) for b, this nonlinearity in b and g is nearly transformed to a plane in Log(g) and Log(b). Approximating this iterated logarithm with a plane is appropriate for small to moderate values of K (say K<100). When K is estimated with historical data, estimates much smaller than 10 (99% of SICs) are the rule. If estimates of K are averaged across SICs, this average is about 1.6, with most SICs much smaller than 1.6 and a very few larger (some much larger, the distribution of the Ks across SICs is skewed). For values of K<.5 say, the three dimensional graph of this iterated logarithm is visually indistinguishable from a plane. From now on refer to the expression for Log(RV(LR)) given by (3.6) as

Figure 1. Graph of Log Relative Variance (WGVF, given by 3.6) as a function of Log(b) and Log(g).



WGVF. Figure 1. shows WGVF for K=1.6 and $\mathrm{Log}(H+\sum_{i\neq k}^{H} r(|i-k|)) = 1.7$ graphed as a function of Log(b) and Log(g).

The surface in (Log(g),Log(b)) given by (3.6) is virtually parallel to the surface: Log(g)-Log(b) (i.e. they have nearly the same first partial derivatives with respect to (Log(g), Log(b)). This fact suggests that a generalized variance function of the form $f_1 + f_2\mathrm{Log}(H) + f_3\mathrm{Log}(g/b)$, where f_3 is close to one, or $f_1 + f_2\mathrm{Log}(H) + \log(g/b)$, be investigated. Such GVFs are considered in the next section.

4. SOME SIMULATION RESULTS

A GVF is derived from historical CES data where Log(RV(LR)) is fitted to a plane in Log(b) and Log(g). Log(RV(LR)) was estimated from the universe data base where for several thousand (SIC)x(State)x(Month) categories, CES sampling and estimation methodology was replicated 50 times to produce 50 independent link relative estimates. These 50 estimates were used to estimate variance, relative variance, and Log(RV(LR)) using the standard formulae for each of the (SIC)x(State)x(Month) categories, These estimates of Log(RV(LR)) were used to compute a least squares fit of $f_1 + f_2\mathrm{Log}(b) + f_3\mathrm{Log}(g)$. The result of this fit was $f_1=0$, $f_2=-0.7$, and $f_3=1.6$.

This surface is quite different from that hypothesized in the last paragraph of the last section. This is probably due to at least two factors. First, the time distance, H , to the benchmark month is not included as an independent variable; the data seems to prefer that f_1 be zero and the coefficients f_2 and f_3 change as H changes. Second, the historical data used to fit the coefficients (fs) is not uniformly scattered over the rectangle of possible values ([-3,0]x[0,15]) in the (Log(g), Log(b)) plane. The estimation cells with large employment estimates have values of log(g) that vary over the full range, $-3<\mathrm{Log}(g)<0$ but in small estimation cells, log(g) is almost always close to zero (This is a roughly L-shaped region in [-3,0]x[0,15]).

For 16 pairs of values for b (rows) and g (columns) Table 1 contains the values of WGVF with $\mathrm{Log}(H+\sum_{i\neq k}^{H} r(|i-k|)) = \mathrm{Log}(H^{.66})$, K=1.6, H=10, and values of the linearized version of this GVF given by:

$$\mathrm{LGVF}=(-0.7)\mathrm{Log}(b) + (1.6)\mathrm{Log}(g) \quad (4.1),$$

Table 1 indicates that WGVF and LGVF can differ to a disturbing degree over the L-shaped region, low FPCs & high b values or FPCs near one and small to moderate b values. In addition. LGVF is not a function of H, a necessary parameter for the variance of the link relative estimator. The R^2 value associated with the LGVF fit is about .8. The R^2 value for the WGVF (squared correlation between estimates of relative variance and the WGVF predicted values) is also 0.8.

As suggested in the last paragraph of the previous section, the estimates of Log(RV(LR)) described in the first paragraph of this section were used to fit: $f_1 + f_2\mathrm{Log}(H) + \log(g/b)$. The result was $f_1 = f_2 = 2/3$. This GVF relating Log(RV(LR)) to (H,b,g) is called CGVF. and it is also included in Table 1. Recall that CGVF is a plane in (Log(g), Log(b)) that is nearly parallel to WGVF. CGVF is a generalization of the expression for the relative variance of the ratio estimator under simple random sampling. Taking exponentials:

$$\mathrm{RV(LR)}=(g/b)\exp(2/3)H^{2/3} = ((1/b)-(1/\mathrm{LR}))\exp(2/3)H^{2/3}.$$

Table. 1 - Comparison of theoretical and data determined Generalized Variance Functions.

Sample Employment	g = 1/8	g = 3/8	g = 5/8	g = 7/8
1000				
WGVF	-6.7	-5.8	-5.3	-5.0
CGVF	-6.8	-5.7	-5.2	-4.8
LGVF	-8.2	-6.4	-5.6	-5.0
46000				
WGVF	-10.8	-9.7	-9.2	-8.8
CGVF	-10.6	-9.5	-9.0	-8.7
LGVF	-10.8	-9.1	-8.3	-7.7
91000				
WGVF	-11.5	-10.4	-9.9	-9.5
CGVF	-11.3	-10.2	-9.7	-9.4
LGVF	-11.3	-9.6	-8.7	-8.2
136000				
WGVF	-11.9	-10.7	-10.3	-9.9
CGVF	-11.7	-10.6	-10.1	-9.8
LGVF	-11.6	-9.8	-9.0	-8.5

As suggested in the last paragraph of the previous section, the estimates of Log(RV(LR)) described in the first paragraph of this section were used to fit: $f_1 + f_2\mathrm{Log}(H) + \log(g/b)$. The result was $f_1 = f_2 = 2/3$. This GVF relating Log(RV(LR)) to (H,b,g) is called CGVF. and it is also included in Table 1. Recall that CGVF is a plane in (Log(g), Log(b)) that is nearly parallel to WGVF. CGVF is a generalization of the expression for the relative variance of the ratio estimator under simple random sampling. Taking exponentials:

$$\mathrm{RV(LR)}=(g/b)\exp(2/3)H^{2/3} = ((1/b)-(1/\mathrm{LR}))\exp(2/3)H^{2/3}.$$

This is analogous to $(1/n)(1-(n/N))S^2$, where b (a measure of sample size) replaces n and LR (a measure of cell size) replaces N. Thus CGVF is appealing both for it's simplicity and for the fact that after a somewhat complicated derivation, a slight twist on a familiar variance expression is obtained!

5. PUBLISHABILITY RULES AND RELATIVE VARIANCE.

One purpose of estimating relative variance is to establish a statistical criterion for publishing state level estimates that is less arbitrary than the current "15/50 Rule". This rule states that an estimate may be published if it is based on a sample size of at least 15 establishments or the total sample employment is at least 50% of the total cell employment. Experience has shown that this rule is often too stringent and that there are many cells where the 15/50 criteria is not met but the estimates are nonetheless sufficiently stable to publish. This publishability rule is too arbitrary. A more flexible criterion based on a GVF as a measure of reliability is being developed. When the estimate of relative variance using CGVF is "sufficiently small", we would publish. A definition of "sufficiently small" is yet to be determined.

6. CONCLUSIONS

Generalized Variance Functions that relate relative variance to sample employment, H (number of months since last benchmark revision), and the finite population correction factor, appear to be adequate predictors of sampling variability. In addition to computational simplicity of the GVFs, the smoothing GVFs induce is an implicit borrowing of strength across estimation cells. The modest increase in bias of the GVF variance estimator for an estimation cell is more than compensated for by variance reduction. as the entries in Table 2 indicate. We are continuing to study refinements of the GVFs outlined here. Of the three possibilities, the generalized variance function that appears best (simple, intuitive, and it parameters were readily estimable) is CGVF:

$$RV = \mathrm{Exp}[(2/3) + (2/3)\mathrm{Log}(H) + \mathrm{Log}((1/b) - (1/LR))] = (1.93)(g/b)H^{(2/3)},$$

where RV is the relative variance of the link relative estimator of total employment in an estimation cell, b is the total employment for the sample units in that cell, and g is the cell's finite population correction factor, $g=1-(b/LR)$, $LR = LR_H = (BM)\prod_{j=1}^{H}\hat{\beta}_j$ is the link relative estimate for the cell.

REFERENCES

Cochran W.G. (1977). Sampling Techniques, third edition, John Wiley & Sons Inc. Page 153.

Gibrat, R. (1930). Une loi des repartitions economiques: l'effect proportionel, Bulletin de Statistique General, France, 19, 469ff.

Gibrat, R. (1931). Les Inegalites Economiques, Paris: Libraire du Recueil Sirey.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions-1, John Wiley & Sons Inc. Page 115.

Madow, L and Madow, W (1978), "On Link Relative Estimators", ASA Proceedings of the Section on Survey Research Methods, 534-539.

Royall, Richard and Cumberland, W.G. (1978), Variance Estimation in Finite Population Sampling, Journal of The American Statistical Association, 351-358.

Royall, Richard and Cumberland, W.G. (1981), An Empirical Study of the Ratio Estimator and Estimators of its Variance, Journal of the American Statistical Association, 66-77

Royall, Richard and Cumberland, W.G. (1981), Reply to Comments on, "An Empirical Study of the Ratio Estimator and Estimators of its Variance", Journal of the American Statistical Association, 87-88.

Royall, Richard. (1981), "Study of the Role of Probability Models in 790 Survey Design and Estimation" , Bureau of Labor Statistics contract report #80-98.

Royall, R.M. and Eberhardt, K.R. (1975), "Variance Estimators for the Ratio Estimator", Sankhya, Ser C,37,
43-52.

Valliant, R. (1992) Smoothing Variance Estimates for Price Indexes Over Time. Journal of Official Statistics, 8, 433-444.

West, S. (1984), "A Comparison of Estimates for the Variance of Regression-Type Estimators in a Finite Population", ASA Proceedings of the Section on Survey Research Methods,

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